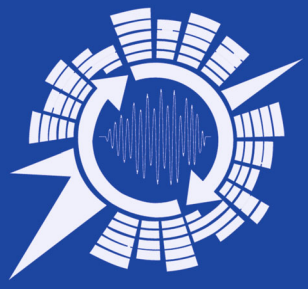
Alexey Karpov
Vlado Delić (Eds.)

LNAI 15299

Speech and Computer

26th International Conference, SPECOM 2024
Belgrade, Serbia, November 25–28, 2024
Proceedings, Part I



Springer

Lecture Notes in Computer Science

Lecture Notes in Artificial Intelligence 15299

Founding Editor

Jörg Siekmann

Series Editors

The series Lecture Notes in Artificial Intelligence (LNAI) was established in 1988 as a topical subseries of LNCS devoted to artificial intelligence.

The series publishes state-of-the-art research results at a high level. As with the LNCS mother series, the mission of the series is to serve the international R & D community by providing an invaluable service, mainly focused on the publication of conference and workshop proceedings and postproceedings.

Alexey Karpov · Vlado Delić
Editors

Speech and Computer

26th International Conference, SPECOM 2024
Belgrade, Serbia, November 25–28, 2024
Proceedings, Part I

Springer

Editors
Alexey Karpov
St. Petersburg Federal Research Center
of the Russian Academy of Sciences
St. Petersburg, Russia

Vlado Delić
University of Novi Sad
Novi Sad, Serbia

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Artificial Intelligence
ISBN 978-3-031-77960-2 ISBN 978-3-031-77961-9 (eBook)
https://doi.org/10.1007/978-3-031-77961-9

LNCS Sublibrary: SL7 – Artificial Intelligence

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

SPECOM 2024 Preface

SPECOM is a conference with a long tradition that attracts researchers in the area of speech technology, including automatic speech recognition and understanding, text–to–speech synthesis, speaker and language recognition, as well as related domains like digital speech processing, natural language processing, text analysis, computational paralinguistics, multi–modal speech, and data processing or human–computer interaction. The SPECOM conference is an ideal platform for know–how exchange – especially for experts working on Slavic languages (e.g. Russian, Serbian, Croatian, Polish, Bulgarian, Czech, etc.) or other inflectional spoken languages – including both under–resourced and regular well–resourced ones.

The International Conference on Speech and Computer (SPECOM) has become a regular event since the first SPECOM, held in St. Petersburg, Russia, in October 1996. The SPECOM conference series was established more than 28 years ago by the St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS).

In its long history, the SPECOM conference was organized alternately by the St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS)/SPIIRAS and by the Moscow State Linguistic University (MSLU) in their home towns. Furthermore, in 1997 it was organized by the Cluj–Napoca subsidiary of the Research Institute for Computer Technique (Romania), in 2005 and 2015 by the University of Patras (in Patras and Athens, Greece), in 2011 by the Kazan Federal University (in Kazan, Russia), in 2013 by the University of West Bohemia (in Pilsen, Czech Republic), in 2014 by the University of Novi Sad (in Novi Sad, Serbia), in 2016 by the Budapest University of Technology and Economics (in Budapest, Hungary), in 2017 by the University of Hertfordshire (in Hatfield, UK), in 2018 by the Leipzig University of Telecommunications (in Leipzig, Germany), in 2019 by the Bogaziçi University (in Istanbul, Turkey), in 2020 and 2021 by SPC RAS/SPIIRAS (fully online), in 2022 by the KIIT (in Gurugram, New Delhi, India), and in 2023 by the IIT/IIIT Dharwad (in Hubli–Dharwad, Karnataka, India).

SPECOM 2024 was the 26th event in the conference series (https://specom2024.ftn.uns.ac.rs), and the second time SPECOM was in the Republic of Serbia. SPECOM 2024 was organized jointly by the Faculty of Technical Sciences at the University of Novi Sad and the School of Electrical Engineering at the University of Belgrade in cooperation with the Telecommunications Society of Serbia. The conference was held between the 25th and 28th November 2024, in a hybrid format, mostly in–person in the capital of Serbia, Belgrade, at the Crowne Plaza Hotel and online via video conferencing. Moreover, SPECOM 2024 was organized jointly and in parallel with the 32nd Telecommunications Forum TELFOR 2024 (https://www.telfor.rs/en). SPECOM 2024 was sponsored and supported by the Science Fund of the Republic of Serbia, as well as by the International Speech Communication Association (ISCA).

During SPECOM 2024, two keynote lectures were given by Dr.–Ing. Kraljevski (Fraunhofer Institute for Ceramic Technologies and Systems IKTS, Dresden, Germany) jointly with his German colleagues on "Preserving Language Heritage Through Speech Technology: The Case of Upper Sorbian", as well as by Prof. Milan Sečujski jointly with his colleagues from the Faculty of Technical Sciences, University of Novi Sad and AlfaNum company, Novi Sad, Serbia on "Retrospective and Perspectives of TTS & STT Technology Development and Implementation for South Slavic Under-Resourced Languages".

This volume contains a collection of submitted papers presented at SPECOM 2024, which were thoroughly reviewed by members of the Program Committee and additional reviewers consisting of over 80 experts in the conference topic areas. In total, 53 regular full papers out of 90 submissions to SPECOM 2024 were carefully selected by the Program Committee members for oral presentation at the conference, as well as for inclusion in these SPECOM 2024 proceedings. Theoretical and more general contributions were presented in common plenary sessions. Problem–oriented sessions as well as panel discussions brought together specialists in niche problem areas with the aim of exchanging knowledge and skills resulting from research projects of all kinds.

We would like to express our gratitude to all authors for providing their papers on time, to the members of the SPECOM 2024 Program Committee for their careful reviews and paper selection, and to the editors and correctors for their hard work in preparing the conference proceedings. Special thanks are due to the members of the SPECOM 2024 Organizing Committee for their tireless effort and enthusiasm during the conference organization. We are also grateful to the Faculty of Technical Sciences at the University of Novi Sad, the School of Electrical Engineering at the University of Belgrade, and the Telecommunications Society of Serbia for organizing and hosting the 26th International Conference on Speech and Computer, SPECOM 2024, in Belgrade.

November 2024

Alexey Karpov
Vlado Delić

Organization

General Chairs

Vlado Delić	University of Novi Sad, Serbia
Alexey Karpov	St. Petersburg Federal Research Center of the Russian Academy of Sciences, Russia

Program Committee

Alexey Karpov (Chair)	St. Petersburg Federal Research Center of the Russian Academy of Sciences, Russia
Vlado Delić (Chair)	University of Novi Sad, Serbia
Shyam Agrawal	KIIT Gurugram, India
Jahangir Alam	Computer Research Institute of Montreal, Canada
Shahin Amiriparian	Technical University of Munich, Germany
Elias Azarov	Belarusian State University of Informatics and Radioelectronics, Belarus
Milana Bojanić	University of Novi Sad, Serbia
Nick Campbell	Trinity College Dublin, Ireland
Vladimir Chuchupal	Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, Russia
Andrea Corradini	Design School Kolding, Denmark
Govind D.	K L University, India
Olivier Deroo	Acapela Group, Belgium
Denis Dresvyanskiy	Ulm University, Germany
Anna Esposito	Università degli Studi della Campania "Luigi Vanvitelli", Italy
Vera Evdokimova	Saint-Petersburg State University, Russia
Nikos Fakotakis	University of Patras, Greece
Mauro Falcone	Fondazione Ugo Bordoni, Italy
Abderrahim Fathan	Computer Research Institute of Montreal, Canada
Olga Frolova	Saint-Petersburg State University, Russia
Jovan Galić	University of Banja Luka, Bosnia and Herzegovina
Suryakanth Gangashetty	KLEF, India
Philip N. Garner	IDIAP Research Institute, Switzerland

Branislav Gerazov	Saints Cyril and Methodius University in Skopje, North Macedonia
Barbara Gili Fivela	Università del Salento, Italy
Gábor Gosztolya	University of Szeged, Hungary
Ivan Gruber	University of West Bohemia, Czech Republic
Denis Ivanko	St. Petersburg Federal Research Center of the Russian Academy of Sciences, Russia
Nikša Jakovljević	University of Novi Sad, Serbia
Ildar Kagirov	St. Petersburg Federal Research Center of the Russian Academy of Sciences, Russia
Alexey Kashevnik	St. Petersburg Federal Research Center of the Russian Academy of Sciences, Russia
Heysem Kaya	Utrecht University, The Netherlands
Maria Khokhlova	Saint-Petersburg State University, Russia
Irina Kipyatkova	St. Petersburg Federal Research Center of the Russian Academy of Sciences, Russia
Liliya Komalova	Moscow State Linguistic University, Russia
Evgeny Kostyuchenko	Tomsk State University of Control Systems and Radioelectronics, Russia
Yanxiong Li	South China University of Technology, China
Natalia Loukachevitch	Moscow State University, Russia
Elena Lyakso	Saint-Petersburg State University, Russia
Ilya Makarov	Artificial Intelligence Research Institute, Russia
Olesia Makhnytkina	ITMO University, Russia
Maxim Markitantov	St. Petersburg Federal Research Center of the Russian Academy of Sciences, Russia
Konstantin Markov	University of Aizu, Japan
Yuri Matveev	ITMO University, Russia
Peter Mihajlik	Budapest University of Technology and Economics, Hungary
Nikolay Mikhaylovskiy	Tomsk State University, Russia
Bernd Möbius	Saarland University, Germany
Sebastian Möller	Technical University Berlin, Germany
Ruban Nersisson	Vellore Institute of Technology University, India
Aleksandar Nešković	University of Belgrade, Serbia
Tijana Nosek	University of Novi Sad, Serbia
Dariya Novokhrestova	Tomsk State University of Control Systems and Radioelectronics, Russia
Sergey Novoselov	STC-Innovations Ltd., Russia
Nick A. Petrovsky	Belarusian State University of Informatics and Radioelectronics, Belarus
Lidia Pivovarova	University of Helsinki, Finland
Branislav Popović	University of Novi Sad, Serbia

Vsevolod Potapov	Lomonosov Moscow State University, Russia
Rodmonga Potapova	Moscow State Linguistic University, Russia
Sergey Rybin	ITMO University, Russia
Dmitry Ryumin	St. Petersburg Federal Research Center of the Russian Academy of Sciences, Russia
Elena Ryumina	St. Petersburg Federal Research Center of the Russian Academy of Sciences, Russia
Milan Sečujski	University of Novi Sad, Serbia
Tatiana Sherstinova	HSE University, St. Petersburg, Russia
Nickolay Shmyrev	Alpha Cephei Inc., Russia
Vasiliki Simaki	Lancaster University, UK
Nikola Simić	University of Novi Sad, Serbia
Pavel Skrelin	Saint-Petersburg State University, Russia
Tatiana Sokoreva	Moscow State Linguistic University, Russia
Victor Sorokin	Institute for Information Transmission Problems of the Russian Academy of Sciences, Moscow, Russia
Ajay Srinivasamurthy	Amazon Alexa, India
Siniša Suzić	University of Novi Sad, Serbia
Jianhua Tao	Institute of Automation, Chinese Academy of Sciences, China
Ivan Tashev	Microsoft, USA
Natalia Tomashenko	University of Avignon, France
Laszlo Toth	University of Szeged, Hungary
Isabel Trancoso	INESC-ID/IST, University of Lisbon, Portugal
Jan Trmal	Johns Hopkins University, USA
Liliya Tsirulnik	Stenograph LLC, USA
Alena Velichko	St. Petersburg Federal Research Center of the Russian Academy of Sciences, Russia
Vasilisa Verkhodanova	University of Groningen, Campus Fryslan, The Netherlands
Zeynep Yucel	Okayama University, Japan
Miloš Železný	University of West Bohemia, Czech Republic
Jerneja Žganec Gros	Alpineon, Slovenia

Additional Reviewers

Nikolay Bobrov
Lidija Krstanović
Bin Liu
Danila Mamontov
Yong Ren
Vuk Stanojev
Anton Stepikhov

Organizing Committee

Vlado Delić (Chair)	University of Novi Sad, Serbia
Milan Sečujski	University of Novi Sad, Serbia
Branislav Popović	University of Novi Sad, Serbia
Milana Bojanić	University of Novi Sad, Serbia
Nikola Simić	University of Novi Sad, Serbia
Nikša Jakovljević	University of Novi Sad, Serbia
Siniša Suzić	University of Novi Sad, Serbia
Tijana Nosek	University of Novi Sad, Serbia
Vuk Stanojev	University of Novi Sad, Serbia
Mladen Koprivica	University of Belgrade, Serbia
Jelena Ćertić	University of Belgrade, Serbia
Alexey Karpov	SPC RAS, Russia
Dmitry Ryumin	SPC RAS, Russia
Irina Kipyatkova	SPC RAS, Russia
Ildar Kagirov	SPC RAS, Russia
Alexandr Axyonov	SPC RAS, Russia

Contents – Part I

Contents – Part II

Speaker Recognition

Digital Speech Processing

Natural Language Processing

Invited Papers

Preserving Language Heritage Through Speech Technology: The Case of Upper Sorbian

Ivan Kraljevski[1(✉)], Frank Duckhorn[1], Daniel Sobe[2], Constanze Tschoepe[1], and Matthias Wolff[3]

[1] Fraunhofer Institute for Ceramic Technologies and Systems IKTS, Dresden, Germany
{ivan.kraljevski,frank.duckhorn,constanze.tschoepe}@ikts.fraunhofer.de
[2] Foundation for the Sorbian People, Bautzen, Germany
d.zoba@zalozba.de
[3] Chair of Communications Engineering, Brandenburg University of Technology (BTU) Cottbus-Senftenberg, Cottbus, Germany
matthias.wolff@b-tu.de

Abstract. The modern world is facing a crisis with the rapid disappearance of endangered languages, which poses a serious threat to global cultural diversity. Speech Technologies and Artificial Intelligence present promising opportunities to address this crisis by supporting the documentation, revitalization, and everyday use of these vulnerable languages. However, despite recent and remarkable advancements in speech technology, significant challenges persist, particularly for languages with very limited resources and unique linguistic features.

This paper details the development of Upper Sorbian speech technologies, focusing on the creation of a practical Speech-to-Text (STT) system as a versatile tool for language preservation. The study explores the current state of Sorbian languages and underscores collaborative efforts with the Foundation for the Sorbian People. Through a series of pilot and successive projects, each phase has contributed to the steady advancement of speech recognition modules and supporting tools, improving their performance, effectiveness and practical usability.

Keywords: Endangered languages · Speech recognition · Upper Sorbian

1 Introduction

In recent times, humanity is experiencing one of the most serious crises: the disappearance of endangered languages. An endangered language is one predicted to cease being used for communication within a specific cultural or social group. This crisis has a profound human dimension, as a language dies together with its last native speaker.

A. Karpov and V. Delić (Eds.): SPECOM 2024, LNAI 15299, pp. 3–22, 2025.
https://doi.org/10.1007/978-3-031-77961-9_1

According to UNESCO's Atlas of the World's Languages in Danger [19], of the over 6,500 languages spoken in the world, more than 2,400 are in danger of disappearing. The majority of the world's population, 97%, speaks only 4% of all languages. Moreover, more than 1,500 languages have fewer than 1,000 speakers. It is expected that more than 90% of the currently spoken languages will disappear by the year 2100, replaced by dominant regional and national languages [27].

The loss of a language has multifaceted and profound implications on cultural, social, linguistic, and psychological aspects. It results in the loss of heritage, cultural and community identity, oral traditions, connections with newer generations, and inter-generational knowledge transfer [6,20].

Advances in information and media technology, particularly Speech Technologies (ST) and Artificial Intelligence (AI), offer new ways to preserve language diversity by aiding traditional methods that rely on written, audio, and visual documents. These technologies can support the documentation, revitalization, and daily use of endangered languages. In daily use, ST helps speakers hear and learn correct pronunciations (Text-to-Speech, TTS) and create written records from spoken language (STT), also aiding in the creation of educational materials.

Recent state-of-the-art ST has made breakthroughs in recognition, achieving "near-human" performance in restricted conditions, domains, and languages. However, the challenges of introducing state-of-the-art STT for a new language are multifaceted, especially if it has limited electronic resources.

If sufficient data for a target language exists or can be collected, the data requirements for reliable speech and language modeling using end-to-end (E2E) systems and deep learning (DL) would be feasible. However, despite recent advances in STT technology for endangered languages, the unique phonetic features and limited language resources make training models and achieving accurate recognition a challenging task, further complicated by dialectical variations, background noise, or non-standard speech utterances.

In this paper, we present the development of the Upper Sorbian speech technologies, focused on a Speech-to-Text system in the context of heritage and language preservation. The objective is not only to create a practical speech recognition system but also to lay a foundation for tools, speech and language resources in Upper Sorbian to be used in wider research fields. Such as, computational phonetics and linguistics, where the experiences and knowledge can be transferred also to the related Lower Sorbian language.

2 Current State of the Sorbian Languages

The Sorbian languages (Upper and Lower) belong to the West Slavic branch of the Indo-European language family, along with Polish, Czech, and Slovak, and are recognized by the European Charter for Regional or Minority Languages [7].

Upper Sorbian (language code: hsb) is spoken in the region of Upper Lusatia, covering parts of the German federal states of Saxony and Brandenburg, while Lower Sorbian (language code: dsb) is spoken in the region of Lower Lusatia, located entirely in southern Brandenburg. Although closely related, these

languages are only partially mutually intelligible, and each branch has retained some distinct features from Old Slavic. In general, Upper Sorbian is perceived to be closer to Czech and Slovak, whereas Lower Sorbian is considered closer to Polish.

Reliable figures on the number of Sorbian language speakers are lacking, with estimates ranging widely. Upper Sorbian is estimated to have between 15 and 30 thousands speakers, while Lower Sorbian has between 5 and 10 thousands speakers [14]. According to [25], the speaker population is estimated to be between 20 and 30 thousands with one-third in Lower Lusatia and two-thirds in Upper Lusatia, out of a total ethnic population of 45 thousands.

Most native speakers do not use the language in daily communication, threatening its existence despite its protected status. Native speakers of Upper Sorbian are a minority in predominantly German-speaking areas [13]. Both languages face significant preservation challenges and are classified as vulnerable or severely endangered by UNESCO, with most fluent speakers being elderly and younger generations having limited proficiency.

Bilingualism with German is common, and despite legal protections and support from umbrella organizations for Sorbian associations in Upper and Lower Lusatia, like "Domowina", the decline in speakers due to socio-political factors poses a threat.

3 Speech Technologies for Upper Sorbian

Substantial efforts have been made to preserve the Sorbian languages through educational programs, media presence, and community initiatives. Notable projects have been conducted by the Foundation for the Sorbian People[1] and the Sorbian Institute[2], with significant contributions from other research groups focusing on the collection and documentation of Sorbian speech and language, particularly emphasizing the digitization of written, audio, and video records.

In [28], the history and development of the Upper Sorbian Textual Corpus (HoTKo) was detailed. This corpus, which includes journalistic, literary, religious, and scientific texts from approximately the mid-nineteenth century to the present, currently contains about 44 million tokens (common word forms). Cooperation with the Domowina Publishing House and the WITAJ[3] Language Center facilitated the inclusion of a vast collection of contemporary texts for research.

The Corpus for Spoken Lower Sorbian (GENIE) [18] compiles audio recordings from the Sorbian Broadcast Archive (1956–2006), the Archive of Sorbian Culture (1951–1971), and new recordings from native speakers (2005–2006). The authors focused on the unique situation of Lower Sorbian and its bilingual speakers, addressing the challenges associated with creating corpora for endangered languages.

[1] https://stiftung.sorben.com.
[2] https://www.serbski-institut.de.
[3] https://www.witaj-sprachzentrum.de.

Between 2010 and 2015, an audio corpus comprising 100 hours of recordings was collected at the Sorbian Institute in Cottbus [3], aiming to document the speech of native Lower Sorbian speakers. Orthographic transcriptions and German translations were provided, with selected samples also transcribed phonetically and translated into English.

The SobLexx project[4], conducted by the Sorbian Institute, the Foundation for the Sorbian People, and the WITAJ Language Center, offers an electronic Upper Sorbian-German dictionary with grammar, spell checker, and word segmentation modules, all freely available online.

In 2021, the first Upper Sorbian-German machine translation (MT) application (SOTRA[5]) was launched with the support of the WITAJ Language Center in Bautzen and Cottbus.

The development of a Text-to-Speech system based on the MARY-TTS framework for Sorbian languages is described in [26].

However, as of 2020, to the best of our knowledge, there were no Upper Sorbian speech corpora suitable for the research and development of speech technologies, specifically speech recognition. The contemporary Common Voice HSB dataset was inadequate due to its limited data and imbalanced speaker distribution.

The cooperation between the "Cognitive Material Diagnostic" (KogMatD) project group at Fraunhofer IKTS, the Chair of Communication Technologies at Brandenburg University of Technology in Cottbus, and the Foundation for the Sorbian People in Bautzen began in early 2020 with a pilot study. The study aimed to investigate the feasibility of creating a Speech-to-Text system using limited resources.

After successful completion of the study, the stakeholders proceeded with four successive projects to develop speech resources and tools for Sorbian languages, with a focus on speech recognition in Upper Sorbian. The goal was to develop and gradually enhance the technologies for the main components of a traditional STT system: acoustic, lexicon, language model and speech recognition engine.

At the beginning of the cooperation, various End-to-End (E2E) systems were available for fine-tuning under-resourced languages, such as Mozilla DeepSpeech [11], Wav2Letter [5], and later Wav2Vec [2] from FAIR (Facebook AI Research), as well as Google Conformer and others. Despite this, we initially chose to follow the traditional approach of developing separate acoustic, lexicon, and language models. This decision was driven by the desire to create speech and language resources that could be applied in other areas of speech technology, such as Text-to-Speech (TTS) or basic research in phonology.

While fine-tuning large pre-trained E2E models offers potential benefits, it also introduces challenges. Issues that arise during fine-tuning cannot be easily isolated and addressed, especially when time and resources are constrained, as was the case in our feasibility study. This limitation led us to prioritize a modular,

[4] https://soblexx.de.
[5] https://sotra.app.

traditional STT approach, where individual components can be independently optimized and troubleshot.

The state of the art in speech-to-text advanced with the introduction of OpenAI Whisper [24] and the availability of pre-trained models in languages similar to Upper Sorbian, such as Czech and Polish. For the first time, Upper Sorbian was included in Meta's (formerly Facebook) Fairseq Massively Multilingual Speech (MMS) model, "MMS-1B-all" [23]. Like OpenAI Whisper and similar approaches, this model outputs sequences of graphemes.

The training of this model included the Upper Sorbian dataset from Common Voice (CV) [1], which is notably imbalanced—one speaker accounts for over 80% of the corpus sentences. While the model generates acoustically plausible outputs with a respectable Character Error Rate (CER), it also exhibits a high Word Error Rate (WER). Consequently, achieving accurate STT necessitates additional language modeling, which is further challenged by the scarcity of textual resources required to train a Large Vocabulary Continuous Speech Recognition (LVCSR) system.

As an illustration, consider the recognition of a sentence recorded in real conditions during a church service. The errors are highlighted in boldface:

Reference: *bojach so a tuž* ***woteńdźech*** *a* ***schowach*** *swój talent* ***do zemje hlej tu maš štož je twoje ale jeho*** *knjez jemu wotmołwi* ***zły a lěni wotročko*** *sy wědźał* ***zo žněju*** *hdźež njejsym syła zběram hdźež njejsym sypał.*

Hypothesis: *bojach so a tuž* ***wotendźek*** *a* ***skowach*** *swój talent* ***dozem lej tumaš štojše twoj ale j ho*** *knjez jemu wotmołwi* ***zwy aleni wotwočko*** *sy wědźał* ***zo dźneju*** *hdźež njejsym sył a zběram hdźež njejsym sypał.* While the model captured the general structure and many correct words, it still struggled with specific phonetic and morphological details of the language.

One significant advantage of our approach is that all performance enhancements are traceable and explainable. This allows us to systematically identify and correct issues within individual modules, ensuring a more, flexible, reliable and maintainable system.

4 Feasibility Study

The main objective of the pilot study was to investigate the feasibility of developing a speech recognizer for Upper Sorbian using existing resources and tools in German. The open-source framework *dLabPro* for signal processing and the *Unified Approach to Signal Synthesis and Recognition (UASR)* frameworks [12] were employed, primarily due to the familiarity with and availability of acoustic models in German. These models can be utilized "out-of-the-box" in forced-alignment for phonetic annotations of collected Upper Sorbian speech recordings.

The practical demonstrator was envisioned as a voice application, specifically a prototypical STT system for a limited language domain-home automation-referred to as the "Smart Lamp" demonstrator. A brief overview is provided in the following sections; more details of the feasibility study can be found in the corresponding paper [16].

4.1 Grapheme and Phoneme Inventory

The first step was to define the grapheme and phoneme inventories for German and Upper Sorbian to facilitate mapping. The phoneme inventory was adopted from the UASR framework, since we use the available acoustic model in German, while the Upper Sorbian inventory was derived from publicly available sources and those provided by the project partner, The Foundation for the Sorbian People.

Public sources included the Wikipedia page for Upper Sorbian[6], the "HSB" language site of the Sorbian Institute[7], and the book excerpt "Obersorbisch im Selbststudium/Hornjoserbšćina za samostudij" (Lesson 02, pages 12–13) [29].

After establishing the grapheme inventory and converting phonemes into X-SAMPA format, the Upper Sorbian phonemes were mapped to their nearest German equivalents. By defining pronunciation rules, we were able to map the graphemes to corresponding phoneme sequences, which is a prerequisite for forced-alignment of the speech and subsequent adaptation of the German acoustic model.

4.2 Pronunciation Rules

The mappings from graphemes to X-SAMPA phonemes are simple "one-to-one" rules, whereas the mappings to the UASR phoneme set include some "one-to-many" and "many-to-one" rules. Pronunciation variants of grapheme sequences within words are defined by exception rules, derived from [29] and further refined with a feedback from native speakers.

The grapheme context (Left_GRPH_Right) specifies the pronunciation of phoneme(s) or their omission, "#C" denotes a consonant, "#V" denotes a vowel, "$" indicates a word boundary, and "*" represents phoneme omission.

For example, in the word "*zymskich*"(/z/ /Y/ /m/ /s/ /k/ /i:/ /C/), the rule "I_CH_" maps to /C/ is applied, whereas in "*zwučowanjach*"(/z/ /U/ /v/ /u:/ /t/ /S/ /O/ /U/ /v/ /a/ /n/ /j/ /a/ /x/), the default grapheme-to-phoneme rule applies, where the digraph "CH" maps to default phoneme /x/.

4.3 Speech Application Specification

One of the main objectives of the study was to provide a practical demonstration of a speech application in a limited domain (voice control). To achieve this, an ontology for the domain of voice control of a smart lamp was developed. Realistic utterance examples were defined using templates that include intents (such as switching on/off, setting brightness, and setting color) along with their parameters, as well as more natural phrases (e.g., "please" - "*prošu*", "dear lamp" - "*luba swěca*").

[6] https://en.wikipedia.org/wiki/Upper_Sorbian_language.
[7] https://www.obersorbisch.de.

The specification was then transformed into a Backus-Naur Form (BNF) grammar, which was used to randomly generate a variety of in-domain sentences. These generated texts served as prompts for speech collection and used also for language modeling.

4.4 Speech Data

The speech data originate from two distinct sources: the validated portion of the Common Voice dataset, which is crowd-sourced and open-source, and a speech corpus collected during controlled recording sessions (HSB speech corpus). Using the defined pronunciation rules, we generated a lexicon that was used to obtain phoneme transcriptions, a prerequisite for the forced-alignment stage in acoustic adaptation.

The **Common Voice hsb-dataset (v5.1)** contains 1,600 audio files from 2 female and 15 male speakers, with a total duration of 2:42:02. The content is sourced from various general-domain materials such as newspapers, books, and proverbs. Sentences containing graphemes and words of foreign origin were omitted, resulting in 1,352 sentences with 5,579 vocabulary entries. The lexicon was reviewed by a native speaker, and pronunciation rules were refined accordingly. Unsuitable words were removed, and the lexicon was used to filter out inappropriate sentences to avoid inconsistencies in phoneme modeling.

The **HSB corpus** was recorded at the premises of the Foundation for the Sorbian People in Bautzen, Germany. Three different phonetically balanced sets of prompts were utilized in the recording sessions. Approximately two-thirds of the textual prompts were selected from the Common Voice (CV) dataset, while the remaining were composed of domain-specific generated sentences for the "Smart Lamp" (SL) application.

4.5 Prompts Selection

To achieve maximum coverage of the phonemic units in controlled recording sessions, the prompts were designed to be phonetically balanced and rich, aligning with the phoneme unit statistics of larger textual data. For this purpose, we utilized the corpus Monolingual Upper Sorbian Data[8], which contains a vocabulary of 251,358 words.

We calculated the frequencies of phones, di-phones, and tri-phones, and selected prompts from both domain-general (CV) and domain-specific (SL) sentences. The selection process was guided by a scoring algorithm applied to di-phones, as presented in [4] and we generated three distinct sets, totaling 1,200 prompts.

[8] https://www.statmt.org/wmt21/unsup_and_very_low_res.html.

4.6 Speech Recordings

A total of 30 speakers were recruited: 10 of females, males, and children. The child participants were minors attending either higher grades of elementary school or lower grades of high school, indicating good reading skills.

Each speaker was instructed to read the prompts exactly as presented, allowing us to target specific pronunciation variants and reducing the need for extensive manual post-processing and transcription of the recordings.

In the end, we collected approximately 11 hours and 30 minutes of speech data, achieving a nearly equal distribution across the three prompt sets.

4.7 Forced-Alignment Experiments

Phoneme annotations are created by initial forced-alignment using the knowledge-based phoneme mappings. Then phoneme recognition performance is evaluated on the **HSB corpus** to discover the most frequent phoneme confusions.

After the initial evaluation of the confusion matrix, it was evident that there are many confusions across the vowels (e.g., /a/ with /a:/, /aI/, /aU/ ...). Therefore, the first data-driven based optimization was to reduce *baseline* German acoustic model from 43 to 29 phoneme models (*reduced* model). This narrows the choice of phoneme sequences, improving the robustness of the acoustic model and reducing confusions, thereby enhancing phoneme recognition accuracy.

However, the acoustic model's performance remains relatively low. To improve it, the model needed adaptation to the language and acoustic environment (e.g., microphones, interface, room acoustics).

4.8 Acoustic Model Adaptation

The adaptation and evaluation of the acoustic models was performed with "Leave One Group Out" cross-validation (LOGO) strategy. The results are aggregated into one data frame where for each recognized sentence, the speaker's recordings were not part of the model adaptation.

The maximum a-posteriori (MAP) algorithm was used to adapt the means and covariances of the Gaussian distributions using the adaptation portion of the HSB dataset. The absolute improvements in phoneme error rates achieved after adaptation were from 66.86% to 39.98% for the baseline acoustic model, and from 61.7% to 37.9% for the reduced phoneme set model.

4.9 Smart Lamp Voice Application

The language model for the "Smart Lamp" application was written in the form of a set of Finite-State-Grammar (FSG) rules. To identify errors and problematic rules we tested and optimized the grammars on the adapted acoustic models to ensure the best recognition performance. The optimization was mostly directed to discover pronunciation variants and speech rate issues with compound words.

We achieved best recognition performance in the case of reduced and LOGO adapted models. However, to deliver a demonstrator that will be both, speaker independent and robust against varying acoustic environments, we adapted the reduced model exclusively on the Common Voice speech data and tested on the HSB dataset. The increase in the error rate was not drastic and it is expected that the model will perform reliably according to the quality of the audio signal. The demonstrator was implemented and made accessible via a webpage, allowing interested users to control a virtual smart lamp.

5 Upper Sorbian Speech Recognition

After successfully concluding the feasibility study, the cooperation progressed to developing a Large Vocabulary Continuous Speech Recognition (LVCSR) application in Upper Sorbian. We utilized conventional HMM/GMM systems, optimizing each underlying model separately. This approach leveraged existing technology (dLabPro/UASR) and facilitated the development from mono-phone to hybrid TDNN/HMM acoustic models, which are compatible with other Speech-to-Text (STT) frameworks.

In the first phase, we used the collected data to perform acoustic training based on Upper Sorbian's native phoneme inventory, rather than relying on models for other languages. Significant advancements were achieved in language modeling by introducing word-class models as Finite-State Transducers (FST), which could be incorporated into Context-Free Grammars (CFG) or Statistical Language Models (SLM).

The second phase aimed to develop resources and tools for domain-specific large vocabulary applications, with plans for future domain-independent use.

In the third phase, the focus shifted to improving and optimizing technologies by collecting and organizing additional speech and language data. With the newly collected resources, the corpus, along with augmented speech and text, expanded to over 70 hours. This expanded corpus was used to train new, more robust acoustic models with both the original and a reduced phoneme inventory.

5.1 Acoustic Modeling

Phase I. Firstly, to accurately capture the phonetic characteristics of Upper Sorbian, pronunciation was modeled using the native phoneme set rather than a subset of German phonemes.

We then utilized the German acoustic model adapted for Upper Sorbian from the feasibility study, where the phoneme models were mapped to the native phonemes. This adapted model was used to perform forced alignment of the **HSB corpus**. As a result, the speech data was aligned with corresponding native phoneme annotations, including precise timing information. This annotated corpus was subsequently used to train a new acoustic model with the dLabPro/UASR toolkit.

Additionally, we introduced speaker-dependent adaptation to enhance the model's performance for individual speakers by fine-tuning it to their unique speech patterns.

Finally, we evaluated and compared the speech recognition performance of both speaker-independent and speaker-dependent acoustic models, observing significant improvements with the speaker-dependent model.

Phase II. To advance from mono-phone based acoustic models to tri-phone based models, it is crucial to collect more transcribed data of spontaneous speech across diverse domains.

For the training of tri-phone acoustic models, we combined speech data from three different sources: the **Common Voice hsb-dataset (v5.1)**, data collected during the feasibility study (**HSB corpus**), and new audio data provided within this project (**SCF corpus**, Speech Corpus Film). The resulting corpus includes 106 speakers, 15,000 recordings, and a total duration of approximately 18 hours.

Clean speech recordings were augmented by introducing various types of background noise at random levels, effectively doubling the amount of available speech data for training. All speech data was augmented except for the Common Voice dataset, which was reserved for evaluating recognition performance. With augmentation, the total duration of the speech recordings increased to around 25 hours, comprising a total of 29,162 utterances.

We used the KALDI open-source STT toolkit [22] to train mono- and tri-phone models with our custom "UPFA" (Universal Primary Feature Analysis) features. Initially, the speech corpus, audio data, and transliterations were converted into KALDI configuration data. Training was conducted for models with varying numbers of states and senones.

The augmented speech corpus was divided into three datasets, ensuring that no speaker appeared in both the training and test/dev sets simultaneously:

- **train** with 24052 recordings, including augmented ones,
- **test** with 3760, similar recording conditions (including augmented utterances),
- **dev** with 1350, is the Common Voice dataset, representing real-use case.

The development set was used only for additional validation, due to the Common Voice speech data nature - crowd-sourced from speakers with different audio equipment, acoustic environments and compressed with the MP3 encoder. We trained and evaluated different tri-phone models, that can be converted back into the dLabPro/UASR format and used with our custom speech recognition engine (recIKTS).

The lexicon and language models are created from the transliterations of training, development, and the test set. The language model is a statistic 3-gram model created with IRSTLM toolkit [10]. Therefore, the results illustrate an ideal case, where the speech is matching the language model.

It is notable also, that the combined transliterations make a corpus of different domains, Common Voice, "Smart Lamp" and captions from movies and

documentaries (Speech Corpus Film). This is not an optimal case, but it gives the impression of the acoustic model quality and performance.

The achieved WER results showed that the tri-phone models have significantly better performance (6.91% [6.10, 7.73] and 7.94% [7.08, 8.80]) than the mono-phone model 10.40% [9.07, 11.72] on the validation set, the Common Voice Data.

As a result, the acoustic modeling procedure employing tri-phones was established and the trained models achieved a level of robustness ensuring speaker-independent recognition in real and adverse conditions.

Phase III. The existing corpus was further expanded with newly provided recordings and transliterations. All the original recordings were augmented, and the resulting corpus duration reached over 70 hours.

In total 58,008 recordings were included in the train, development, and test set:

- Training (131 speakers, 62.32 hours, 52895 recordings) including augmented versions,
- Development corpus (cross-validation) (17 speakers, 2.28 hours) CV common voice dataset v.5.1,
- Test (8 speakers, 6.63 hours).

Additionally, recordings collected from YouTube videos of Sunday church services for domain-specific performance evaluation (unknown speakers, 3.3 hours).

Two phoneme inventories we employed for training of new acoustic models, the *default* one as it was in the previous projects and *reduced* phoneme inventory created by collapsing the vowels; /e/→/E/, /o/→/O/, /u/→/U/.

By collapsing similar phoneme variations into fewer categories, the model becomes less complex and more efficient without sacrificing word recognition accuracy.

The model with fewer phonemes (reduced inventory) achieved better results (WER 3.48%) on the development dataset (the Common Voice corpus) not seen in the training, indicating a more robust acoustic model on unseen data, while the default one achieved better results (WER 1.66%) on the test data which is similar with the training data. As in the previous phase, the language model used for evaluation was created from the combined corpus of the train, development and test sets, although the results are too optimistic, they indicate relative performance improvements.

5.2 Lexicon Modeling

Lexical modeling was performed using previously established basic Grapheme-to-Phoneme (G2P) rules, which were enhanced with an list of exception rules. The lexicon for the given corpus was automatically generated and used for both acoustic modeling and recognition.

Phase I. At this stage, we received a manually created lexicon in Upper Sorbian containing around 4,000 words, which was used as training data for statistical pronunciation modeling. The statistical modeling was conducted using the Python wrapper[9] for the grapheme-to-phoneme tool Phonetisaurus [21]. This tool utilizes N-gram based translation models and is typically implemented as a weighted finite state transducer (WFST).

Given the size of the training data, its suitability for reliable statistical modeling with WFSTs was uncertain. We trained the model and performed a basic comparative analysis of both rule-based and statistical G2P modeling approaches. The resulting model can generate pronunciations for words unseen in the training data, with the option to provide the N-best candidates.

We found out that around 25% of the automatically generated pronunciations using the rule based G2P mappings are matching the ones manually created by a phonetician. The differences are mostly due to the incompatible phoneme inventories (Table 1).

Table 1. Phonetician vs. knowledge-based G2P.

Word	IPA	SAMPA	UASR-HSB	Match
ŠOŁ	/ʃo^{h}/	S o h	S o	No
ZAWRJENKA	/zaurɛŋka/	z a u r E n k a	z a u r E n k a	Yes
BĚHANIŠĆO	/bejaniʃtʃɔ/	b e i a n i S tS O	b ji h a n i S tS O	No
NAHNIĆ	/naniʧ/	n a n i tS	n a n i tS	Yes
NAHRABAĆ	/narabatʃ/	n a r a b a tS	n a r a b a tS	Yes

Phase II. The exception rules are further enhanced by defining some of them as mandatory or optional. The mandatory rules are always applied and replace the pronunciations generated by simple grapheme-to-phoneme mappings (canonical), while the optional rules add pronunciation variants to the lexicon.

The generator script is further improved to include handcrafted lexicons. Words with existing pronunciations are not processed further and are simply added to the resulting lexicon.

This approach allows for the handling of words with foreign origins, including German words and proper names that can often be declined according to Upper Sorbian grammatical cases. For instance, "Sam běch na tekst storčił, jako mějach nadawk, Försterowu knihu ...," where the proper name "Förstero + wu" is declined.

Therefore, foreign language graphemes were integrated into the grapheme inventory (such as 'ä', 'ö', 'ü', 'ñ', etc.), and suitable phonemes from the previously defined phoneme inventory were assigned.

Phase III. To further enhance the lexicon modeling, we investigated the following aspects:

[9] https://github.com/rhasspy/phonetisaurus-pypi.

Redefinition of Symbols. We redefined symbols in the inventory, particularly the phonemes "jn" and "ji", after comparing them with the pronunciation dictionary used for Upper Sorbian language synthesis.

Comparative Evaluation. We assessed the models for the sounds "e", "o", and "jn", and, if necessary, integrated them into existing models as "E", "O", and "n".

Phoneme Inventory Reduction. After comparative evaluation of the phoneme inventory, we determined that the position of articulation (open, closed) for some vowels does not significantly impact speech recognition as it does in speech synthesis. Therefore, we reduced the default phoneme inventory by mapping vowels as follows: /e/→/E/, /o/→/O/, and /u/→/U/. Significant modifications to the phoneme inventory were not considered in relation to the "MARY-TTS-HSB" project, as its objectives differ substantially.

Listeners of Upper Sorbian can easily detect minor mispronunciations in a Text-To-Speech (TTS) system. However, in speech recognition, these nuances are less frequent and statistically less significant for acoustic modeling. The pronunciations from the "MARY-TTS-HSB" project were qualitatively compared with those automatically generated using the "default" phoneme inventory. It was concluded that, overall, there are no major differences in pronunciation for the Speech-to-Text (STT) task.

5.3 Language Modeling

Speech recognition for highly inflected languages (such as Upper Sorbian) poses challenges for language modeling due to the many word forms that must be included in the vocabulary.

We investigated and developed tools and procedures for text processing and normalization, word-class modeling with recognition of named entities (NER), and tokenization in sub-word units (e.g., morphemes).

Phase I. The concept of CFG grammars used in the "Smart Lamp" demonstrator were extended to be used in modeling of statistical language models with word classes.

Combining CFG grammars with SLMs allow significantly more freedom in expression. Terms like numbers, time, date, places, proper nouns were specified with rule-based grammars for each word class. These and other word classes can be integrated into CFG grammars as well as in a statistical language models. We give an overview of the word-class modeling, the detailed description is given in [17].

We developed CFG grammars as extended weighted finite-state-transducers (FST) [8] that convert Upper Sorbian number, time, date, currency and percentages expressions into a proper numerical representation.

Each edge of an FST takes a sub-string of an expression as input and converts it into an arithmetical operation consisting only of addition and multiplication

operators. The FSTs can be also nested, incorporating smaller ones that represent more elementary expressions.

The advantage of using FST based grammars is that they can be seamlessly used for word-class modeling by including them in FST decoding graph, and at same time, as grammars to parse textual content and perform Named Entity Recognition (NER).

The parser is enhanced with regex rules that match the cardinal and ordinal numbers, time, date, percentage and currency.

Numbers. We developed a basic numeral grammar noted as NUM1-9 in the OpenFst-TextFile Format, consisting of the cardinal numerals from 1 to 9. The basic grammar is incorporated to construct the grammar NUM1-99. Smaller grammars are combined to recursively construct larger ones like NUM1-99, NUM1-999, NUM1-10^6, and so on.

Time of Day. These grammars convert time of day expressions into a numerical representation of the time. We do not use the classical $hh : mm$ format but rather the count of minutes after midnight.

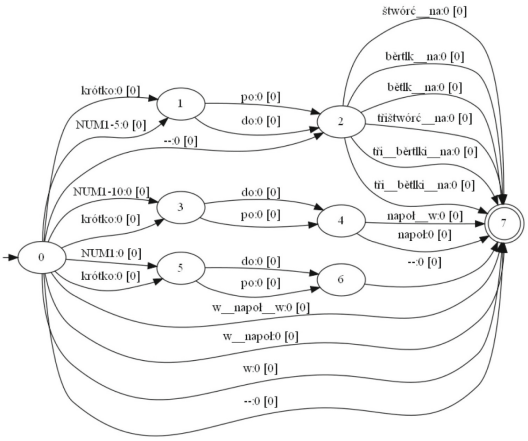


Fig. 1. Sorbian hour time modifiers. Note: NUMz-y represents a sub-grammar FST for the numerals from z to y.

We are considering two different types of time expressions. One that covers the accurate digital expressions that can be simply modeled out of a numeral between $0-23$ as the hour count, a numeral between $0-59$ as the minute count and "hodźin" as a connection word between the hour and minute count. The second type covers the more common - but complicated - everyday expressions like "tři štwórć na pjećich" (corresponding to "quarter to five") (Fig. 1).

Date. For the date grammars, we again decided to represent the numerical meaning in a single number - the day count after New Year's Eve and mostly use negative numbers. Since there are leap years, the day count after New Year's

Eve is indefinite for any date from March to December. However, the day count till New Year's Eve is definite, so we use negative counts for March till December but positive counts for January and February. So, the output is always a number between −305 and 60. Moreover, we built two different date grammars for nominative and genitive case. Both cases are needed for some significant wordings.

The date grammars are built out of an ordinal number representing the day and a name for the month. We included 3 different names for each month: A numerical name as the ordinal number of the month, a Gregorian name and an older traditional name. In the FST model we combine the month names with the ordinal number grammars ORD1-29, ORD1-30, ORD1-31 or ORD1-31f, depending on the length and gender of the month (name).

Phase II. The size of a vocabulary is a problem for the language modeling, particularly in very inflected languages, the size of the lexicon could reach several million word-forms. Then the rate of out-of-vocabulary (OOV) words is very high which makes reliable recognition not feasible.

One of the solutions is to use sub-word modeling, where the words are broken down to smaller units (tokens), effectively reducing the size of the vocabulary. There are many approaches that are used for tokenization words in NLP system and as pre-processing stage for training of Large Language Models (LLMs) and here we employed the Byte-Pair-Encoding (BPE) and the Morfessor algorithms.

Text Corpus Processing. We developed procedures to process and normalize electronic language resources provided in various formats. These procedures organize, import, and convert the resources into suitable formats. They also preprocess and normalize the data, preparing it for subsequent processing stages and ultimately for language modeling.

For normalization the following workflow is developed:

- Import, processing, and normalization of text in different formats (such as, doc, pdf, html, xml).
- Normalization of abbreviations, replacement with tags or spoken words, removing redundant punctuation and special characters.
- Named Entities Recognition (NER) from defined word classes (names, time, date, numbers, locations), substitution of word classes in the normalized corpus and creation of corresponding FST grammars.
- Segmentation of words into tokens (e.g., morphological units) by an automatic parser and checking of the segmentation based on the pronunciation of the words and units.
- Generation of a vocabulary and a lexicon for the sub-word units.

Performance Evaluation. We evaluated the performance of the language models over text data of the domain of interest - transcriptions of the Sunday's divine services in the Parish Church in Chróśćicy (Crostwitz), Saxony. Additional textual resources collected from different sources were used for training

(unsupervised and supervised) of the BPE and Morfessor tokenizer models. The chosen metric for evaluation are the relative differences in the corpus size increase (less is better), unique token decrease (more is better) and average token length decrease (less is better).

The N-gram language models are trained on the 95% of the randomly shuffled sentences and we calculate the perplexity.

The perplexity was used as the criteria to choose the parameters for the LM model and the tokenization approach, after that all the available textual data was used to train the final language model and build the recognizer engine configuration.

Phase III. This phase targets the creation of domain-specific language models based on the already collected and normalized texts, with additional new domain-specific texts and their normalization and integration. Possible improvements were also investigated in the already adopted approach that uses tokenized morphological units, also considering other natural language processing (NLP) tokenization algorithms.

The application domain was narrowed to closed captioning of audio recordings of church services ("Boža mša z Chróśćic") with possible use in live broadcasting and offline transcriptions to provide compatible subtitles for YouTube. Live broadcasting is challenging due to the adverse acoustic environment. In this case, there are also periods of non-speech events, such as organ music and choir performing. Additionally, the textual data from the domain is still scarce, albeit most of the speech is related to the religious texts, there are somewhere future events that are announced and they contain proper names of persons, locations, time, and date expressions. The domain-specific corpus contains verified and normalized transcripts of the recorded church services collected over a longer period.

The amount of in-domain text is still very small, therefore, the source corpus contains also out-of-domain texts. However, sub-word modeling ensures unseen word contexts are to some extent covered.

It is important to emphasize that the main issue is still non-matching corpus with unseen recordings. The solution is continuous improvement of the language models by transcribing as many as possible broadcasts. We increased the amount of texts by randomly swapping word places in sentences in two iterations with a ratio of 1:2 between the original and the augmented texts. The final 4-gram language model was created by interpolating the models different domain previously processed into sub-words with the Wordpiece tokenizer.

While testing the recognition on new unseen recordings, it was noticeable that the acoustic conditions differ significantly compared to the original recordings. The new unseen recordings have a much higher level of reverberation (echo). This renders the recognition unfeasible for live transcriptions, even offline captioning is difficult due to the poor recognition performance. Therefore the corpus was augmented with echoic recordings simulation the room impulse response. The "Room" dimensions and the absorption characterization were given as much as similar to the location where the recordings are actually taken.

The procedure was successfully applied for rapid transcription of the church services and closed-captioning of the Youtube broadcasts. In general, when the sub-word contexts are seen in the training data, the recognition is robust accurate, and practically usable.

5.4 Speech Recognition Engine

Fraunhofer IKTS has developed proprietary software for speech and signal recognition (recIKTS), designed to offer a broader range of features compared to freely available alternatives, with the added benefit of potential commercial applications. Written in C and C++, the software is available as both a stand-alone application and a library. It is compatible with various architectures and operating environments, including Win-32/64, Linux-i386/amd64/arm64, and is optimized for signal processors.

This software handles the entire pipeline, from audio input to feature extraction, I-vector calculation, acoustic model computation, and decoding. A key focus of the implementation is resource efficiency. While speech recognition is a primary use case, the software is versatile enough to recognize technical and biological signals as well.

For feature extraction, the software supports classic MFCC features, including LDA feature transformation, which are compatible with Kaldi's MFCC features. Alternatively, it offers customizable Fourier, wavelet, or cepstrum transformations for technical signals, which can be paired with a configurable filter bank and principal component analysis. I-vector calculations for TDNN models are also supported and can be performed in real-time during signal input.

The acoustic models used in the software are trained externally and imported into the IKTS recognizer. The software supports TDNN and HMM models from Kaldi for mono-, bi-, and tri-phones, as well as UASR HMM mono-phone models.

Unlike Kaldi, the language model in this software offers greater flexibility. It supports statistical language models in ARPA format, predefined CFG grammars, or precompiled automata in OpenFST format. The language model can also incorporate word classes, which can be defined as word lists, grammars, or OpenFST automata. Sub-word modeling is also supported, allowing the language model to include sub-word units or morphemes that are reassembled during post-processing.

The core component of the speech recognition engine is its highly optimized decoder, which has been engineered for maximum resource efficiency. It is built on an efficient FST implementation and features a custom variant of the token-passing algorithm [9]. This decoder supports live detection for recordings of any length, with recognition results provided in real-time. The software uses an iterative backtracking algorithm to refine these results.

In addition to recognition capabilities, the software includes speech pause detection. For this purpose, it integrates the GMM-based VAD from the open-source WebRTC project. A dedicated state machine manages trigger offsets and minimum activation and deactivation times.

6 Conclusion

We presented the development of speech technologies for Upper Sorbian, focusing on speech recognition in collaboration with the Foundation for the Sorbian People.

The development process was structured around the core components of a traditional speech recognition system: acoustic, lexicon, and language modeling. The project was executed in three phases, each corresponding to short-term objectives. These efforts significantly contributed to the preservation of the Upper Sorbian language by creating valuable tools, electronic language and speech resources.

The speech corpus was expanded significantly from its initial size, which comprised a relatively small number of sentences collected in controlled recording sessions. This expansion involved adding more transcribed and aligned speech from various sources, including recordings, movies, and documentaries, culminating in over seventy hours of augmented data. The resulting corpus and the associated experiences are valuable assets for research in linguistics and phonetics, as well as for various speech applications.

Our collaboration with the Foundation's representatives during the development of the lexicon, word-class, and language models was instrumental. This partnership led to the discovery of lesser-studied language features and raised new scientific questions. For example, we conducted a data-driven study on the occurrence of glottal stops before word-initial vowels in Upper Sorbian, examining the effects of speaker demographics (males, females, and children) and vowel types [15].

The corpus and lexicon provided a foundation for the continuous development and training of mono-, bi-, and tri-phone GMM/HMM, and TDNN/HMM acoustic models, achieving robust and reliable recognition performance.

The availability of these core components, combined with domain-specific language modeling, paves the way for numerous applications of speech recognition. These include new speech data collection, transcription of spoken documents for preservation, human-machine interaction, accessibility tools for disabled individuals, education, and language learning.

In a practical application, we configured and utilized the speech recognizer to transcribe divine church service broadcasts, a challenging task due to the acoustic environment and limited textual data. This procedure significantly accelerated the creation, editing, and publishing of transcripts.

Recent technological advances in speech technology cannot be overlooked. We are actively working on new developments and creating synergies between the latest state-of-the-art STT systems and the resources we have developed. These efforts aim to further improve and popularize speech technologies in Upper Sorbian, contributing to the preservation of both the language and its cultural heritage.

Acknowledgments. This study was supported by the Foundation for the Sorbian People in Bautzen, Germany. The authors would like to thank Mr. Jan Budar, Mrs.

Michaela Moosche, Mrs. Katharina Čornak, and Mr. Christian Richter for their support and assistance. We acknowledge the use of ChatGPT (OpenAI, 2024) for language proofing of the manuscript.

References

1. Ardila, R., et al.: Common voice: a massively-multilingual speech corpus. In: Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pp. 4211–4215 (2020)
2. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: a framework for self-supervised learning of speech representations. In: Advances in Neural Information Processing Systems, vol. 33, 12449–12460 (2020)
3. Bartels, H., Thorquindt-Stumpf, K.: Ein neues Ton-und Textarchiv des muttersprachlich-dialektalen Niedersorbischen (2013)
4. Berry, J., Fadiga, L., et al.: Data-driven design of a sentence list for an articulatory speech corpus. In: INTERSPEECH, pp. 1287–1291 (2013)
5. Collobert, R., Puhrsch, C., Synnaeve, G.: Wav2letter: an end-to-end convnet-based speech recognition system. arXiv preprint arXiv:1609.03193 (2016)
6. Crystal, D.: Language Death. Cambridge University Press, Canto (2002)
7. Dołowy-Rybińska, N.: A model minority. ACADEMIA. Mag. Pol. Acad. Sci. **Nr 2 (30) 2011 Fear**, 48–49 (2011)
8. Duckhorn, F., Hoffmann, R.: Using context-free grammars for embedded speech recognition with weighted finite-state transducers. In: Proceedings of the 13th Annual Conference of the International Speech Communication Association (Interspeech 2012), pp. 1003–1006. Portland, OR, USA (2012)
9. Duckhorn, F., Wolff, M., Hoffmann, R.: A new epsilon filter for efficient composition of weighted finite-state transducers. In: Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech 2011), pp. 897–900. Florence, Italy (2011). https://doi.org/10.1109/EUROCON.2013.6625203
10. Federico, M., Bertoldi, N., Cettolo, M.: IRSTLM: an open source toolkit for handling large scale language models. In: Interspeech (2008). https://api.semanticscholar.org/CorpusID:34745880
11. Hannun, A., et al.: Deep speech: scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567 (2014)
12. Hoffmann, R., Eichner, M., Wolff, M.: Analysis of verbal and nonverbal acoustic signals with the Dresden UASR System. In: Esposito, A., Faundez-Zanuy, M., Keller, E., Marinaro, M. (eds.) Verbal and Nonverbal Communication Behaviours. LNCS (LNAI), vol. 4775, pp. 200–218. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-76442-7_18
13. Howson, P.: Upper Sorbian. J. Int. Phon. Assoc. **47**(3), 359–367 (2017)
14. Jodlbauer, R.: Die aktuelle Situation der niedersorbischen Sprache : Ergebnisse einer soziolinguistischen Untersuchung der Jahre 1993–1995 (2001)
15. Kraljevski, I., Bissiri, M.P., Duckhorn, F., Tschoepe, C., Wolff, M., et al.: Glottal stops in upper Sorbian: a data-driven approach. In: Interspeech, pp. 1001–1005 (2021)
16. Kraljevski, I., Rjelka, M., Duckhorn, F., Tschöpe, C., Wolff, M.: Cross-lingual acoustic modeling in upper Sorbian - preliminary study. In: Hillmann, S., Weiss, B., Michael, T., Möller, S. (eds.) Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2021, pp. 43–50. TUDpress, Dresden (Mar (2021)

17. Maier, I., Kuhn, J., Duckhorn, F., Kraljevski, I., Sobe, D., Wolff, M., Tschöpe, C.: Word class based language modeling: a case of Upper Sorbian. In: Ojha, A.K., Ahmadi, S., Liu, C.H., McCrae, J.P. (eds.) Proceedings of the Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia within the 13th Language Resources and Evaluation Conference, pp. 28–35. European Language Resources Association, Marseille, France (2022). https://aclanthology.org/2022.eurali-1.5
18. Marti, R., Andreeva, B., Barry, W.: GENIE: The corpus for spoken lower Sorbian (GEsprochenes NIEdersorbisch). Phonetician **101/102**, 47–59 (2010)
19. Moseley, C.: Atlas of the World's Languages in Danger. Unesco (2010)
20. Nettle, D., Romaine, S.: Vanishing Voices: The Extinction of the World's Languages. Anthropology online, Oxford University Press (2000). https://books.google.de/books?id=UKNhAAAAMAAJ
21. Novak, J.R., Minematsu, N., Hirose, K.: Phonetisaurus: exploring grapheme-to-phoneme conversion with joint n-gram models in the WFST framework. Nat. Lang. Eng. **22**(6), 907–938 (2016). https://doi.org/10.1017/S1351324915000315
22. Povey, D., et al.: The Kaldi speech recognition toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society (2011). IEEE Catalog No.: CFP11SRW-USB
23. Pratap, V., et al.: Scaling speech technology to 1,000+ languages (2023). https://arxiv.org/abs/2305.13516
24. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision (2022). https://arxiv.org/abs/2212.04356
25. Salminen, T.: Europe and North Asia, pp. 211–280. Routledge, International (2007)
26. Schmiedel, A., Steiner, I.: Development of speech syntheses for lower Sorbian and upper Sorbian using Marytts (2023)
27. Turin, M.: Voices of vanishing worlds: endangered languages, orality, and cognition. Análise Soc. **47**(205), 846–869 (2012)
28. Wölkowa, S.: The upper Sorbian text corpus and further sources of information with regard to upper Sorbian in the internet [tekstowy korpus a dalše informaciske srědki wo hornjoserbskej rěči w interneće]. Studia z Filologii Polskiej i Słowiańskiej **49**, 59 (2014). https://doi.org/10.11649/sfps.2014.008
29. Šołćina, J., Wornar, E.: "Obersorbisch im Selbststudium/Hornjoserbšćina za samostudij", Ein Sprachkurs für Unerschrockene (inkl. CD), vol. 3. Aufl., Bautzen. Domowina-Verlag (2012)

Retrospective and Perspectives of TTS & STT Technology Development and Implementation for South Slavic Under-Resourced Languages

Milan Sečujski[1(✉)], Branislav Popović[1], Darko Pekar[2], Nikša Jakovljević[1], Edvin Pakoci[2], Siniša Suzić[1], Tijana Nosek[1], Nikola Simić[1], Vuk Stanojev[1], and Vlado Delić[1]

[1] Faculty of Technical Sciences, University of Novi Sad, Novi Sad, Serbia
secujski@uns.ac.rs
[2] AlfaNum Ltd, Novi Sad, Serbia

Abstract. Speech technologies such as text-to-speech (TTS) and speech-to-text (STT) are becoming increasingly applicable. Significant improvements in their quality are driven by advancements in deep machine learning. The ability of devices to deeply understand human speech and generate appropriate responses is a hallmark of AI capabilities. Developing speech technology requires extensive speech and language resources, which is why many languages with smaller speaker bases lag behind widely spoken languages in the development of speech technologys. Prior to the deep learning (DL) paradigm, hidden Markov models (HMM) and probabilistic approaches dominated speech technology development. This paper reviews the challenges and solutions in TTS and STT development for Serbian, highlighting the transition from HMM to DL. It also explores the future prospects of speech technology development for under-resourced languages and its role in preserving these languages.

Keywords: Speech Technology · Development and Implementation · Text-to-Speech · Speech-to-Text · Hidden Markov Models · Deep Neural Networks

1 Introduction

The development of speech technology, encompassing both text-to-speech (TTS) and speech-to-text (STT) systems, has revolutionized human-computer interaction and significantly impacted numerous fields, including accessibility, communication, and artificial intelligence. While the initial research focus was to address fundamental challenges in signal processing and language modeling, the last few decades have seen remarkable improvements, driven primarily by the emergence of machine learning algorithms and the vast availability of training data. With the advent of statistical methods and, more recently, deep learning (DL) techniques, modern TTS and STT systems have become more robust, adaptive, and capable of learning from large-scale datasets. These systems are now ubiquitous, powering virtual assistants, voice chatbots, transcription services, and accessibility tools that have become integral to everyday life [1].

A. Karpov and V. Delić (Eds.): SPECOM 2024, LNAI 15299, pp. 23–42, 2025.
https://doi.org/10.1007/978-3-031-77961-9_2

The paper explores the historical evolution of speech technology, outlining the key milestones, technical challenges and innovations, with particular focus on under-resourced languages [2]. Namely, while both technologies have made significant strides in dominant languages like English, their application to under-resourced languages has historically lagged. Under-resourced languages often lack large-scale linguistic datasets, making the development of high-quality TTS and STT systems for these languages particularly challenging. The authors of the paper are members of a research team behind the development of first fully functional and commercially widely applied TTS and STT systems for Serbian and several other South Slavic languages, and besides giving a general historical review of speech technology, the authors will focus on the issues and challenges they have encountered in the development of speech technology for under-resourced languages.

The remainder of the paper is structured as follows. In Sect. 2 we will present a historical overview of speech technology, with emphasis on their language-dependent elements. Section 3 will focus on particular challenges encountered in the development of speech technology for Serbian and other kindred languages. Section 4 focuses on the issues related to under-resourced languages, and Sect. 5 will conclude the paper.

2 Evolution of Speech Technology

Technological advances in the field of artificial intelligence and machine learning have been followed by our perpetually changing perspective on speech technology. In their beginnings, both speech recognition and synthesis have been viewed as typical signal processing areas and focused on topics such as speech coding [1]. The development of first commercial TTS or STT systems required specialized knowledge of linguistics, turning speech technology into a prime example of interdisciplinary knowledge area, where the tasks of conversion of text into speech or vice versa are decomposed into smaller subtasks corresponding to different tasks in human speech recognition and production, requiring different knowledge and relying on different types of speech and language resources (speech corpora, text corpora, lexicons, rule lists, statistical models). However, the recent developments in artificial intelligence and machine learning have shifted the focus towards deep learning systems using sophisticated neural network architectures whose components exhibit little correspondence with particular human speech production or recognition tasks. As a result, both TTS ans STT are now viewed as typical instances of machine learning problems, whose success is due to the ability of neural networks to model the complexity of human language, learn from vast amounts of data, and generalize to unseen speech or text inputs. One of the most important advantage of this shift in perspective is the possibility of using transfer learning, which allows pre-trained models (typically trained on a large, resource-rich dataset in a major language) to be fine-tuned to new, often low-resource languages. This approach reduces the need for massive amounts of labelled data, which is often unavailable for under-resourced languages [2 – 4].

2.1 Text-To-Speech Synthesis

The history of text-to-speech (TTS) technology spans several decades, evolving from early mechanical devices to today's sophisticated systems based on artificial intelligence. The first known speech synthesizer, VODER, was developed by Homer Dudley at Bell Labs in 1939 [5]. It could produce basic speech sounds using a keyboard but required manual operation. In 1961 IBM created the *IBM 704*, one of the earliest examples of computer-generated speech, which used formant synthesis to emulate human vocal tract shapes. The first full TTS system for English was introduced in 1968 by Teranishi and Umeda [6]. In the 1980s, digital signal processing advanced TTS with the development of DECtalk, relying on a source-filter algorithm [7], which provided a more natural-sounding synthesized voice.

The most common feature of the first commercial text-to-speech systems, able to convert any text in a given language (in this case English) into speech, was their internal structure, which was divided into parts charged with language processing (referred to as *front end*) and signal processing (referred to as *back end*) [8]. Until quite recently, this division, shown in Fig. 1, has represented the joint feature of all practically applicable text-to-speech architectures.

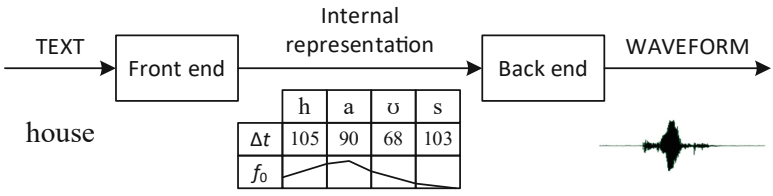


Fig. 1. Internal structure of a classical TTS system.

Front End. The front end firstly converts raw text containing symbols such as numbers and abbreviations into their orthographic equivalents, which is often referred to as *text pre-processing* or *text normalization.* The front end then assigns phonetic transcriptions to each word, which is referred to as *text-to-phoneme* or *grapheme-to-phoneme* conversion. The text is then segmented out into prosodic units such as phrases and sentences. The output of the front end is the linguistic representation of the input text, including its phonetic transcription and prosodic information. The back end, also referred to as the *synthesis module,* subsequently converts the linguistic representation from its symbolic form into sound. In most TTS systems this task, which is practically language independent, includes the computation of the desired prosody features (phone durations, pitch contour), which are subsequently imposed on output speech waveforms [9].

Humans are able to perform tasks related to text normalization, grapheme-to-phoneme conversion and prosody generation automatically, owing to inference capabilities of their brains. In doing so, humans unconsciously exploit their entire linguistic competence, which is most notable in the task of prosody generation. Namely, although prosody is predominantly affected by lexis and syntax, it also appeals to higher levels of linguistic competence of the reader, including semantics and pragmatics. In machines,

the modules for prosody generation are unavoidably simplistic, and are usually further segmented into morphological analysis, contextual analysis and syntactic-prosodic parsing. For most languages, the morphological analyzer proposes all possible part-of-speech (POS) categories for each word, the contextual analyzer considers each word in its context and reduces the list of its possible POS categories to a very small number of highly probable hypotheses, while the syntactic-prosodic parser examines the remaining search space relating it to the expected prosodic realization [8]. All front-end tasks are heavily language dependent, which means that a front end for each new language had to be developed anew. Modules for all these tasks could be implemented as either rule based or based on some form of machine learning, but in either case, they required the existence of a speech or language resource whose creation typically required a significant effort of experts (e.g. rule sets or labelled text or speech data). The advent of machine learning and neural networks has introduced many changes into this paradigm. Initially, neural networks were used as the method of choice for particular front-end tasks, most notably grapheme-to-phoneme conversion [10, 11]. They have achieved notable results in prosody generation from linguistic cues identified by the front end [12], and the recent advances in end-to-end TTS are aimed at completely eliminating the need for the front end as a separate module in a TTS pipeline.

Back End. The back end, also referred to as the *synthesis module,* converts the linguistic representation from its symbolic form into sound. The widespread usage of the TTS technology came with the introduction of concatenative synthesizers, with the idea of producing speech by concatenation of prerecorded speech segments. While some of the early systems used a fixed-size unit inventory for synthesis [13, 14], a true improvement in speech quality came with dynamic unit selection from large speech databases [15]. Although this approach, assuming a very large speech database is available, produces high-quality speech, there are still audible glitches at the concatenation points if the appropriate units cannot be found in the database. Furthermore, this approach is also extremely inflexible in terms of changing the speaking style or the voice of the speaker, which can be done only by recording and annotating a new speech database.

With the increasing popularity and demand for TTS, the demand has also grown for algorithms able to produce different voices and speaking styles from small data samples. The turn of the century saw the advent of statistical parametric speech synthesis, based on modelling the spectrum, fundamental frequency, and duration of speech by multispace probability distribution hidden Markov models (HMM) and multidimensional Gaussian distributions [16]. This approach enables the transformation of a speaker-independent system toward a target speaker using very small samples of speech data [17], creating expressive voices [18], as well as multilingual voices [19]. However, this method never achieved the naturalness of concatenative TTS, principally due to smoothness caused by modelling similar contexts with the same Gaussian mixtures, but also to the use of inferior vocoders, i.e. systems that produce speech waveforms from predicted acoustic features. A detailed review of HMM-based TTS can be found in [20]. Some approaches have combined parametric synthesis with unit selection, which is referred to as hybrid synthesis. The most common hybrid systems in general use parametric based models to drive unit selection [21, 22].

Some of the first attempts to use neural networks for TTS are reported in [23]. However, this approach has since gained popularity and eventually taken precedence over other approaches, mostly owing to recent development of computer hardware, particularly graphical processing units (GPUs). Deep neural networks (DNN) replaced decision trees and Gaussian mixture models in non-linear mapping of linguistic features to acoustic features [24]. They also represent a form of parametric synthesis, in that a model is used and trained on a large dataset, inferring values of parameters that will be used to synthesize speech at runtime. Intelligible and natural sounding synthetic speech could be produced even by relatively simple feedforward neural networks, and further improvements were achieved by using long short-term memory (LSTM) neurons [25], generative adversarial networks [26] and stacked bottleneck features [27].

Advanced TTS Features and Approaches. Deep neural networks have also introduced advanced possibilities such as flexible synthesis in different voices and speaking styles. Most methods for creating new voices using limited amounts of training data are based on multispeaker models, requiring a large database consisting of multiple speakers, with each speaker usually represented with less data than in case of single-speaker models [28]. In such models the variety of contextual information and better network generalization usually yield higher TTS quality. Different modalities for speaker representation have been used, including unique speaker vectors [29, 30] as well as the division of the neural network into parts shared across all speakers and speaker-specific parts [31].

The ability of a TTS to convey different emotional states or styles is a necessity for many applications, since it has been shown that emotion, mood, and sentiment affect attention, memory, performance, judgment, and decision-making in humans [32]. Initial approaches to emotional speech synthesis were focused on statistical modelling of speech parameters with HMMs [33, 34] and Gaussian mixture models [35], while more recent advances exploit deep neural networks [36, 37] and deep bi-directional LSTM (DBLSTM) [38, 39]. Further improvement in performance has been achieved with end-to-end neural network architectures [40, 41], while some of the most recent advances include synthesis of mixed emotions [42].

A significant advance in the quality of DNN TTS came with the WaveNet architecture [43], able to directly predict raw audio samples instead of using a vocoder, relying on predictive distributions dependent on previous audio samples. Conditioned on linguistic features derived from text and speaker identity, it significantly outperforms all other TTS systems, and its drawbacks related to extreme computational complexity were somewhat mitigated by the introduction of approaches such as Parallel WaveNet [44]. A similar model called DeepVoice [45], was based on replacing all parts of TTS pipeline by corresponding independently trained DNNs, but this resulted in a cumulative error in synthesized speech in the end.

As opposed to WaveNet and DeepVoice, which still use some form of front end and generate speech based on lexical features, there are systems which use raw orthographic text as input, such as Tacotron [46], Tacotron 2 [47], and Deep Voice 3 [48]. Tacotron outputs spectrograms that are transformed to speech samples using the Griffin-Lim algorithm, which also introduces artifacts in generated speech. On the other hand, Tacotron

2 system-generated spectrograms are used for conditioning standard WaveNet architecture, which generates speech samples. DeepVoice 3 can output spectrograms or other features which can be used as input to some waveform synthesis models.

Adaptation to new speakers has also been investigated in end-to-end systems [49, 50] as well as synthesis in different styles [40, 51]. Tacotron 2 offers high speech quality but can be slow and prone to issues like word skipping. FastSpeech [52] improves on this by using a Transformer network for faster, parallel mel-spectrogram generation, reducing word skipping and allowing smoother voice speed control. While FastSpeech relies on a complex teacher-student distillation process and suffers from inaccurate duration predictions and information loss, FastSpeech 2 [53] addresses these issues by training directly with ground-truth data and incorporating additional speech variations like pitch and energy, improving training speed and voice quality.

End-to-end systems, while eliminating the need for detailed labeling (such as prosody annotation), require vast amounts of data, which must typically be of high quality and often from the same speaker to achieve high-quality TTS. However, even in these conditions, such systems can struggle with certain aspects. One significant drawback is the lack of control over specific language-dependent features [54] or the exact output, which can lead to unwanted artifacts or distortions known as hallucinations.

One of the advanced approaches for TTS is VALL-E [55], a neural codec language model. Unlike previous methods, which treat TTS as continuous signal regression, VALL-E frames it as a conditional language modeling task. By treating TTS as a sequence generation problem, VALL-E leverages discrete neural audio codecs and a GPT-3-like architecture for its robust performance. Owing to in-context learning, it can synthesize high-quality, personalized speech from just a 3-s speech sample. VALL-E outperforms existing zero-shot TTS in naturalness and speaker similarity.

Extensions include VALL-E-X for cross-lingual zero-shot TTS [56], and VALL-E-R [57], which enhances speech generation robustness with phoneme monotonic alignment. VALL-E 2 further improves performance with repetition-aware sampling and grouped code modeling, achieving human-level parity on LibriSpeech and VCTK datasets. MELLE [58], another approach, generates mel-spectrograms directly from text, bypassing vector quantization. VALL-E offers superior zero-shot TTS performance, speaker adaptation, and control over diverse speech attributes but comes at the cost of higher computational requirements.

Another advanced TTS system is YourTTS [59], a multilingual zero-shot end-to-end TTS. Built on the VITS framework, it allows for accent and style transfer, which means it can synthesize speech with the style of a specific speaker, even in a different language. Its zero-shot learning feature enables it to mimic new voices based on short audio samples. It is usually trained on a large multilingual dataset for multiple speakers and uses neural waveform generation methods such as HFG [60] or WaveGlow [61].

While YourTTS offers broad multilingual capabilities and zero-shot learning for new speakers, VALL-E focuses on high-fidelity voice cloning and adaptation to specific voices. YourTTS is more versatile across languages and accents, whereas VALL-E excels in replicating individual voices with high accuracy.

While many zero-shot multi-speaker TTS systems, like YourTTS and VALL-E-X, are limited to several high-resource languages, the XTTS system addresses this limitation by enabling multilingual training and improving voice cloning [62]. Building on the

Tortoise model, XTTS introduces novel modifications for faster training and inference, and it has been trained in 16 languages, achieving state-of-the-art results in most of them. This advancement significantly broadens the applicability of zero-shot TTS to include low and medium resource languages.

The development of Serbian TTS has kept pace with advancements in modern technology, ensuring that its quality level has always been on par with TTS systems for global languages. As early as, in the 2000s, the first Serbian concatenative TTS was developed at the Faculty of Technical Sciences in Novi Sad [63]. Within the collaboration with the company AlfaNum, founded in 2003, the system was continuously improved and initially applied as a screen reader for the visually impaired. The first HMM-based TTS for Serbian was created in 2012 [64], but its quality was not sufficient to replace the already high-quality concatenative TTS for practical applications. However, in 2017 a DNN-based TTS for Serbian was developed [65], which soon surpassed the quality of concatenative TTS, while also enabling flexibility in multi-speaker synthesis [66]. A further step, achieving synthesis quality nearly indistinguishable from human speech, was made possible with the use of HFG-based vocoders [67]. Today, work continues on further improvements, heading towards end-to-end systems [45].

2.2 Speech-To-Text (Speech Recognition)

The first electrical STT systems, developed in the 1950s and 1960s exploited formant energies to recognize isolated phonemes, syllables and digits [68, 69]. The common approach for the first generation of STT systems exploited knowledge of articulatory and acoustical phonetics.

One of the main issues in these first systems were the variations in the duration of the same acoustic unit, which was overcome in the 1970s with the introduction of dynamic programming [70, 71]. Another advance was that instead of formants, linear predictive coefficients (LPCs) were introduced [72], under the assumption that the vocal tract can be modeled as an all-pole system, which yielded a more precise acoustic representation as in this way the entire spectrum envelope was taken into consideration.

The development of digital electronics in the 1970s and the 1980s shifted the focus towards more complex STT tasks – STT systems were required to recognize entire sentences with vocabularies containing hundreds of different words [73, 74]. The task was split into 3 layers – acoustic, lexicon and grammar/language layer. The acoustic layer connected acoustic representations of phonemes with the phonemes or simpler recognition units. The lexicon layer connected these acoustic units with the words, and grammar/language layer defined possible sequences of words to reduce the complexity of the search space. At the same time, acoustical modelling began to be treated as a problem of sequence decoding in noisy telecommunication channels [75].

This statistical approach based on hidden Markov models (HMMs) [76–78], was the most prevalent approach for acoustic modeling until the early 2010s and the emergence of deep learning. HMM in combination with a Gaussian mixture model (GMM) was an effective way to model time (HMM) and acoustic (GMM) variability of phonemes. To model coarticulation, a context dependent phoneme (i.e. triphone) became a basic modeling unit [79], with triphones spanning 3 HMM states. As increasing the number

of models with a fixed amount of available training data reduces the amount of data for the training of each model, different state tying procedures were proposed [80]. To increase the robustness to noise new features based on human perceptual model were introduced, such as mel-frequency cepstral coefficients (MFCCs) [81] and perceptual linear predictive coefficients (PLP) [82]. Since HMM in combination with GMM is a generative model, in order to reduce in-class variability different normalization methods were introduced. Various cepstral mean and variance normalization techniques were introduced to reduce the channel variability in case of MFCC [83–85] and PLP [86]. One of the reasons for the longevity of HMM-GMM is the efficient introduction of discriminative training criteria in model training (maximum mutual information [87, 88], minimum classification error [89] and minimum phoneme error [90]). However, discriminative models gain accuracy if the number of observations per parameter is sufficiently large [91, 92].

The knowledge of the relationships between words and their phonetic transcriptions is typically stored in lexicons, which are usually created manually. Initial language models represented manually created graphs allowing limited numbers of possible word sequences. As the number of possible words rises, it becomes impractical to create such models manually and statistical *n*-gram models were introduced [93, 94]. *N*-gram models calculate scores proportional to the probabilities of *n*-word-long sequences based on texts from newspapers, books and other documents rather than spontaneous language. Different methods have been applied to achieve *n*-gram smoothing [95, 96].

Recent years brought a significant shift in paradigm – statistical based systems have been replaced with systems based on artificial neural networks, or more precisely, deep neural networks (DNN). Although there were successful experiments with STT based on neural networks in the 1980s and the 1990s [97, 98], their low speed was a significant problem. The first paper reporting comparable performance between neural network based STT systems and conventional ones was [99], and 3 years later DNN were reported to outperform state-of-the-art HMM-GMM systems by a wide margin [100]. A big step towards end-to-end models was made by introducing connectionist temporal classification (CTC), which allows DNN training for a sequence labelling task with unknown input-output alignment [101]. Several years later, end-to-end recurrent neural network (RNN) with beam search reached the performance of benchmark systems [102], eliminating much of the complex infrastructure of modern STT systems. State-of-the-art systems of today are based on transformers [103, 104]. The introduction of transformers trained in a semi-supervised manner has overcome problems related to training STT models for low-resource languages [105].

Although self-supervised models such as wav2vec [106] or wav2vec-S [105] can learn speech representations, they require adaptation for specific tasks such as STT. On the other hand, OpenAI Whisper [107] demonstrated the ability to perform STT (and tasks such as language recognition or translation) without additional fine-tuning, but with a requirement for 680,000 h of multilingual data. Whisper supports 99 languages, but with a huge disproportion in the amount of data for each language (65% of training data is English), which is reflected in higher WER for low-resource languages. Initial efforts in fine-tuning Whisper to Serbian, based on large existing datasets for Serbian and Croatian [108], are described in the following chapter.

3 Implementation Challenges and First Applications in Serbian

In its treatment of the issue of under-resourced languages in the development of speech technology, the paper focuses specifically on Serbian and related South Slavic languages. The early development has been driven by a collaborative research effort between the Faculty of Technical Sciences in Novi Sad, Serbia, and the company AlfaNum. As the only team consistently working on speech technology in the region, they had to create the first speech and language resources for several South Slavic languages, develop tools for speech annotation, design application programming interfaces (APIs), and provide support for various operating systems and platforms.

AlfaNum TTS [109] is a leading text-to-speech synthesis system that offers versions in Serbian, Croatian, Bosnian, and Montenegrin, incorporating natural intonation elements. The system delivers near-human voice quality through built-in intonation and accentuation features, significantly enhancing the naturalness of the generated speech. Additionally, it allows for adaptation to a specific speaker's voice with minimal speech data and can generate expressive speech for various applications.

AlfaNum STT [110] is an advanced continuous speech recognition system designed for Serbian, Bosnian, Croatian, and Montenegrin. Specialized language and acoustic models are employed as part of leading regional cloud-based and on-premise automatic speech recognition solutions, including commercial applications for medical and legal dictation, as well as voice assistant mobile applications.

As will be presented in more detail in the following sections, a wide range of applications of AlfaNum's speech technologies – both TTS and STT – have already been developed and deployed in Serbia or elsewhere in the region (Fig. 2).

3.1 TTS Applications

The first TTS application developed using AlfaNum TTS for people with disabilities was anReader, developed for the visually impaired [111], enabling them to use computers and smartphones equipped with screen readers such as JAWS or NVDA. To facilitate the use of anReader, it was necessary to develop a speech API. AnReader is officially recognized as an assistive tool for the visually impaired in Serbia, but its use has extended throughout the western Balkans.

AlfaNum TTS aids individuals with dyslexia by enhancing their reading speed and supports those with congenital or acquired speech disorders. Individuals with speech impairments can type their intended messages, which TTS can then vocalize. Owing to voice conversion, laryngectomized users can use speech synthesizers to replicate their own voices using only several minutes of their earlier speech recordings.

Augmentative alternative communication (AAC) aids support those with limitations in producing or comprehending spoken or written language, including conditions such as cerebral palsy, autism, and intellectual disability. AAC devices range from simple aids, such as picture boards for requesting food or assistance, to sophisticated speech-generating devices. A notable multilingual AAC application, cBoard, employs AlfaNum TTS for several South Slavic languages.

The most basic commercial applications of AlfaNum TTS are used for voice announcements in public transportation. In these applications, TTS with remote access

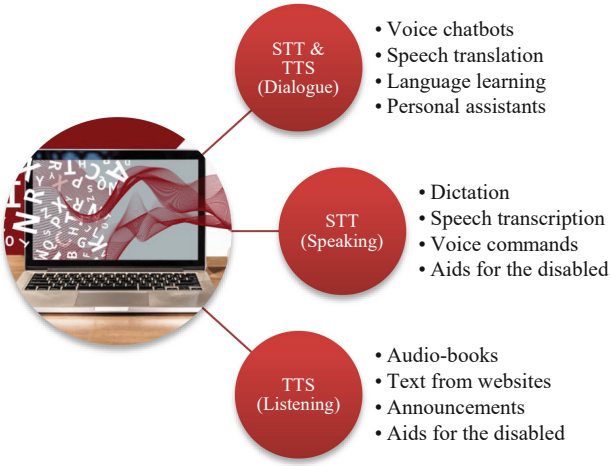


Fig. 2. TTS and STT applications developed and deployed in Serbia and neighboring countries.

(either cloud or on-premise) can be used, as well as the MS-SAPI5 interface if the main application is written for the Windows operating system.

On the other hand, the most commonly used TTS application is the web service "*Read me*", enabling users to listen to news articles in the background while performing other tasks. This feature is widely used on the websites of public media services and various government and public institutions in the region. Some media services, such as the Radio Television of Serbia, even use cloned voices of their own presenters. Implementation of such services faces additional challenges, since AlfaNum provides TTS functionality but does not have access to the internal organization of the web site, which is usually handled by another company. The common practice is to provide high-level libraries for accessing TTS (PHP, Java, Python), and implement the service in collaboration with this company.

The first audio library for the visually impaired was established at the Library of the Union of the Blind in Belgrade. It operates as a client-server system, allowing visually impaired users to access a large database of books via a local network or the internet. The system provides audible output without the need for a separate screen reader, and enables navigation through chapters, paragraphs, and bookmarks. To protect copyright, books are encrypted on the client side and can only be accessed in the audio format. Such a service is also beneficial for individuals who cannot hold books due to physical disabilities, but has become increasingly popular among those who simply prefer audiobooks. The Audio Library of the University of Novi Sad was also developed based on the same idea, and there is also fruitful collaboration with publishers of textbooks for elementary and high school education in Serbia based on TTS [112].

3.2 STT Applications

Even a small vocabulary STT system integrated into smart home technologies can significantly enhance accessibility for the disabled by allowing them to control devices such

as lights and appliances through voice commands. On the other hand, large vocabulary STT systems are suitable for speech transcription and online dictation. There are many potential users of such services, including media outlets, medical or legal practitioners, government agencies etc. GPU-based computers can transcribe speech several times faster than real-time. Some of the existing STT solutions for Serbian, developed within the collaboration of the Faculty of Technical Sciences and AlfaNum are specifically tailored for on-premise users, such as medical and legal institutions.

The MEDICTA and IURISDICTA systems are advanced dictation tools designed to enhance the efficiency of medical and legal professionals by converting dictated speech into text [113]. MEDICTA is tailored for medical findings and operates in real-time on standard computers with regular microphones. It accurately interprets acronyms, punctuation, and initial capital letters, and even allows code-switching between Latin and Serbian. In contrast, IURISDICTA is specifically designed for legal document dictation, effectively recognizing legal terminology and acronyms. Both systems achieve WER below 2%, support user-defined commands, and allow for efficient manual correction of misrecognized words. They also support the use of templates to streamline the dictation of frequently repeated sections, further increasing efficiency.

TRANSCRIPTA is an advanced transcription system that converts recorded speech into text using the open-source Whisper model [107]. It generates transcripts from various audio sources, including TV shows, meetings, conferences, and court hearings. It accurately recognizes natural speech from multiple speakers (WER below 10%, CER below 5%). This level of accuracy was achieved by fine-tuning Whisper with datasets developed for Serbian and Croatian, allowing the system to transcribe Serbian with remarkable precision. TRANSCRIPTA also incorporates a diarization option that distinguishes between different speakers in the audio. Combined with time markers embedded in the transcripts, this enables quick searches through audio and video archives and enhances the efficiency of listening and manual correction of transcripts.

3.3 STT&TTS Applications

Joint applications of STT and TTS facilitate two-way human-machine communication. The development and implementation of these systems in the western Balkans are still in the early stages. AlfaNum's first personal assistant, Axon Voice Assistant, was created for mobile phones, allowing users to make calls, send messages, and perform voice dialing of contacts, addressing complex morphology of Serbian names [114]. However, adaptation to a variety of phone models and operating systems posed a significant challenge for a small company such as AlfaNum, which is why the system was never commercially released. Personal voice assistants are now being integrated not only into smartphones but also into robots, smart speakers, and smart home systems. The future of speech technology in personal assistants looks promising, as advancements in AI enable machines to not just recognize words but also to identify speakers and interpret their moods and intentions.

AlfaNum's TTS systems for Serbian, Montenegrin, Bosnian, and Croatian are currently being integrated into a mobile speech translator that supports over 60 languages. The process of aligning protocols and APIs necessary for accessing AlfaNum's STT and

TTS components is underway, as well as ensuring high throughput during peak times and minimal response latency.

Finally, following the remarkable progress of chatbots like ChatGPT, the next steps involve developing voice chatbots that use STT and TTS, alongside NLP. They will offer functionalities similar to those of call centers, as most calls will be managed automatically by chatbots, either in place of or in conjunction with a smaller number of human operators. We are currently at a stage where providers of end-to-end voice chatbot solutions are expanding into Serbia and other countries where AlfaNum offers advanced TTS and STT capabilities. Again, supporting standard APIs, high throughput and low latency is crucial for high-quality voice chatbots. For TTS, it is usually expected to have a latency of less than 0.5 s, while for STT, it can be somewhat higher since the system requires the entire user's query in order to respond, rather than just the first word. TTS is expected to support multiple speakers and styles, while STT is adaptable to a specific dictionary and language model that best suits the user's needs.

4 Paradigm Shifts in the Development and Perspectives of Speech Technology for Under-Resourced Languages

Advancements in artificial intelligence and natural language processing have profoundly influenced our interactions with technology. TTS and STT systems are among the most prominent technologies that have emerged from these advancements. Although the implementation and evolution of TTS and STT technology have been rapid for many widely spoken languages, the adoption and effectiveness of these technologies face considerable challenges when addressing under-resourced languages [2, 115]. This section examines unique challenges encountered in the deployment of TTS and STT systems for such languages, again, taking Serbian as an illustrative example.

For widely spoken languages, TTS and STT technologies have undergone extensive development and integration into various applications. These languages benefit from the existence of large and diverse datasets necessary to train high-performance DNN-based STT and TTS systems. For instance, TTS systems in widely spoken languages are tapically able to produce voices with various accents, regional dialects, and speaking styles [116], while under-resourced languages face various challenges that impede even the basic functionality of TTS [117] and STT [118] systems.

Open-source initiatives provide essential resources and tools for developing TTS and STT systems for under-resourced languages, promoting collaboration and innovation by granting open access to technology and data for the industrial and academic community [46, 107, 119]. The emergence of open-source solutions has guided local companies toward developing products for specialized domains. By leveraging efficient and adaptable open-source solutions, local stakeholders can create products tailored to specific user needs. This approach reduces costs and allows for customization but also benefits from community support, accelerating development and ensuring they remain competitive and relevant. The development of speech technology for under-resourced languages was significantly facilitated by the use of transfer learning [120]. By adapting large pre-trained models, the existing general knowledge can be leveraged and the models tuned for a specific language or its regional variant to enhance the performance

of TTS and STT while reducing the need for extensive new datasets. Multilingual models are particularly useful, providing dialectical variation by training on corpora from several different languages.

Involving local experts and stakeholders in the development process enhances the accuracy and relevance of TTS and STT technology. This approach ensures that the technology is tailored to local dialects, cultural preferences, and specific user needs, leading to more effective and broadly adopted TTS and STT solutions. In the case of the Serbian language, the collaboration between the Faculty of Technical Sciences in Novi Sad and the company Alfanum resulted in the development of a diverse range of speech resources and speech technology applications for the Serbian language [121]. Initially, the production of these resources required a significant amount of manual labeling, which was labor-intensive and time-consuming. As the project advanced, the adoption of state-of-the-art technologies enabled automatic transcription, significantly reducing the need for expert supervision. This transition, combined with the emergence of publicly available tools for developing speech models, accelerated the development and improved the scalability and efficiency of creating and updating language resources, allowing more rapid adjustments and refinements, and leading to more robust and comprehensive TTS and STT applications.

5 Conclusion

The paper discussed the paradigm shift in the development of text-to-speech (TTS) and speech-to-text (STT) technologies, highlighting the transition from hidden Markov models (HMMs) to deep learning (DL) models. It also explored future perspectives on speech technology applications for under-resourced languages, offering a historical overview and addressing the specific implementation challenges encountered in developing speech technology for Serbian and kindred South Slavic languages.

The case study of Serbian illustrates not only the challenges and solutions in the development of speech technology for under-resourced languages but also the specifics of implementation and exploitation in limited markets. These topics are analyzed and compared across the HMM and DL paradigms. The shift from HMM to DL has facilitated the development of speech technology for under-resourced languages. However, achieving greater independence from global AI giants requires systematic efforts to create speech and language resources for each language. This is why Serbia has established a National Program for Language Technology Development for Serbian as part of its broader AI development strategy. The program aims to create a comprehensive framework for developing speech recognition and synthesis, natural language processing, and other linguistic technologies. It focuses on resource and application development, research and innovation, and training and education, all intended to significantly enhance the capabilities of speech and language technologies in the region while fostering economic growth as well as cultural preservation.

6 Disclosure of Interests.

The authors have no competing interests to declare that are relevant to the content of this article.

Acknowledgments. This research was supported by the Science Fund of the Republic of Serbia, Grant No. 7449, Multimodal multilingual human-machine speech communication, AI-SPEAK, and by the Ministry of Science, Technological Development and Innovation (Contract No. 451–03-65/2024–03/200156) and the Faculty of Technical Sciences, University of Novi Sad through project "Scientific and Artistic Research Work of Researchers in Teaching and Associate Positions at the Faculty of Technical Sciences, University of Novi Sad" (No. 01–3394/1).

References

1. Delić, V., et al.: Speech technology progress based on new machine learning paradigm. Computational Intelligensce and Neuroscience, Wiley, Article 4368036, 19 pages (2019)
2. Besacier, L., Barnard, E., Karpov, A., Schultz, T.: Automatic speech recognition for under-resourced languages: a survey. Speech Commun. **56**, 85–100 (2014)
3. Swietojanski, P., Ghoshal, A., Renals, S.: Unsupervised crosslingual knowledge transfer in DNN-based LVCSR. In: Workshop SLT, pp. 246–251. IEEE, Miami, FL, USA (2012)
4. Tan, X., Qin, T., Soong, F., Liu, T.Y.: A Survey on Neural Speech Synthesis. arXiv preprint arXiv:2106.15561 (2021)
5. Dutoit, T.: High Quality Text-To-Speech Synthesis of the French Language. Ph.D. dissertation. Supervised by Prof. Henri Leich. Faculté Polytechnique de Mons. (1993)
6. Teranishi R., Umeda N.: Use of pronouncing dictionary in speech synthesis experiments. In: Reports of the Sixth International Congress on Acoustics, vol. 2, pp. 155–158 (1968)
7. Hallahan, W.I.: DECtalk Software: text-to-speech technology and implementation. Digit. Tech. J. **7**(4), 5–19 (1995)
8. Dutoit, T.: An Introduction to Text-to-Speech Synthesis. Kluwer Academic Publishers, Dordrecht, Boston, London (1999)
9. Van Santen, J.: Assignment of segmental duration in text-to-speech synthesis. Comput. Speech Lang. **8**(2), 95–128 (1994)
10. Sejnowski, T., Rosenberg, C.R.: Parallel networks that learn to pronounce English text. Complex Syst.1, 145–168 (1987)
11. McCulloch, N., Bedworth, M., Bridle J.: NETspeak – a re-implementation of NETtalk. Comput. Speech Lang. **2**, 289–301 (1987)
12. Ronanki, S.: Prosody Generation for Text-to-Speech Synthesis. Ph.D. thesis, University of Edinburgh (2019)
13. Sagisaka, Y., Kaiki, N., Iwahashi, N., Mimura, K.: ATR v-TALK speech synthesis system. In: Proceedings of International Conference on Spoken Language Processing, pp. 483–486 (1992)
14. Donovan, R.E., Eide, E.: The IBM trainable speech synthesis system. In: Proceedings of 5th International Conference on Spoken Language Processing (ICSLP 98), p. 4, ISCA, Sydney, Australia (1998)
15. Hunt A.J., Black A.W.: Unit selection in a concatenative speech synthesis system using a large speech database. In: Proceedings of ICASSP, pp. 373–376. IEEE, Atlanta, GA, USA (1996)
16. Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura T.: Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis, In: Proceedings of the 6th EUROSPEECH, pp. 2347–2350. Budapest, Hungary (1999)
17. Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., Isogai J.: Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. IEEE Trans. Audio Speech Lang. Process. **17**(s1), 66–83 (2009)

18. Yamagishi, J., Onishi, K., Masuko, T., Kobayashi, T.: Modeling of various speaking styles and emotions for HMM-based speech synthesis. In: Proceedings of the 10th EUROSPEECH, pp. 2461–2464. Geneva, Switzerland (2003)
19. Qian, Y., Liang, H., Soong, F.K.: A cross-language state sharing and mapping approach to bilingual (Mandarin-English) TTS. IEEE Trans. Audio Speech Lang. Process. **17**(6), 1231–1239 (2009)
20. Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J., Oura, K.: Speech synthesis based on hidden markov models. Proc. IEEE **101**(5), 1234–1252 (2013)
21. Yan, Z.-J., Qian, Y., Soong, F.K.: Rich-context unit selection (RUS) approach to high quality TTS. In: Proceedings of ICASSP, pp. 4798–4801. IEEE (2010)
22. Qian, Y., Soong, F.K., Yan, Z.J.: A unified trajectory tiling approach to high quality speech rendering. IEEE Trans. Audio Speech Lang. Process. **21**(2), 280–290 (2013)
23. Weijters, T., Thole, J.: Speech synthesis with artificial neural networks. In: Proceedings of the IEEE International Conference on Neural Networks, pp. 1764–1769, San Francisco, CA, USA (1993)
24. Zen, H., Senior, A., Schuster, M.: Statistical parametric speech synthesis using deep neural networks. In: Proceedings of the ICASSP, pp. 7962–7966. IEEE (2013)
25. Fan, Y., Qian, Y., Xie, F.L., Soong, F.K.: TTS synthesis with bidirectional LSTM based recurrent neural networks. In: Proceedings of 15th INTERSPEECH, pp. 1964–1968. ISCA, Singapore (2014)
26. Saito, Y., Takamichi, S., Saruwatari, H.: Statistical parametric speech synthesis incorporating generative adversarial networks. IEEE/ACM Trans. Audio Speech Lang. Process. **26**(1), 84–96 (2018)
27. Wu, Z., King, S.: Improving trajectory modelling for DNN-based speech synthesis by using stacked bottleneck features and minimum generation error training. IEEE/ACM Trans. Audio Speech Lang. Process. **24**(7), 1255–1265 (2016)
28. Fan, Y., Qian, Y., Soong, F.K., He, L.: Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis. In: Proceedings of ICASSP, pp. 4475–4479. IEEE (2015)
29. Wu, Z., Swietojanski, P., Veaux, C., Renals, S., King, S.: A study of speaker adaptation for DNN-based speech synthesis. In: Proceedings of the 16th INTERSPEECH, pp. 879–883, Dresden (2015)
30. Hojo, N., Ijima, Y., Mizuno, H.: An investigation of DNN-based speech synthesis using speaker codes. In: Proceedings of the 17th INTERSPEECH 2016, pp. 2278–2282. San Francisco, USA (2016)
31. Fan, Y., Qian, Y., Soong, F.K., He, L.: Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis. In: Proceedings of ICASSP, pp. 4475–4479. Brisbane, Australia (2015)
32. Brave, S., Nass, C.: Emotion in human-computer interaction. In: Sears, A., Jacko, J.A. (eds.) Human-Computer Interaction Fundamentals, pp. 53–68, CRC, Boca Raton, USA (2009)
33. Yamagishi, J., Onishi, K., Masuko, T., Kobayashi, T.: Modeling of various speaking styles and emotions for HMM-based speech synthesis. In: 8th EUROSPEECH, Geneva, Switzerland (2003)
34. Eyben, F., et al.: Unsupervised clustering of emotion and voice styles for expressive TTS. In: Proceedings of ICASSP, pp. 4009–4012. IEEE (2012)
35. Aihara, R., Takashima, R., Takiguchi, T., Ariki, Y.: GMM-based emotional voice conversion using spectrum and prosody features. Am. J. Signal Process. **2**(5), 134–138 (2012)
36. Lorenzo-Trueba, J., Henter, G.E., Takaki, S., Yamagishi, J., Morino, Y., Ochiai, Y.: Investigating different representations for modeling and controlling multiple emotions in DNN-based speech synthesis. Speech Commun. **99**, 135–143 (2018)

37. Luo, Z., Chen, J., Takiguchi, T., Ariki, Y.: Emotional voice conversion with adaptive scales F0 based on wavelet transform using limited amount of emotional data. In: Proceedings of the 18th INTERSPEECH, pp. 3399–3403. ISCA (2017)
38. Ming, H., Huang, D., Xie, L., Wu, J., Dong, M., Li, H.: Deep bidirectional LSTM modeling of timbre and prosody for emotional voice conversion. In: Proceedings of the 17th INTERSPEECH 2016, pp. 2453–2457. ISCA (2016)
39. An, S., Ling, Z., Dai, L.: Emotional statistical parametric speech synthesis using LSTM-RNNS. In: Asia-Pacific Signal and Information Processing Association Annual Samit and Conference (APSIPA ASC), pp. 1613–1616, IEEE (2017)
40. Skerry-Ryan, R., et al.: Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron. In: Proceedings of the 34th International Conference on Machine Learning, pp. 4693–4702. PMLR (2018)
41. Wu, P., Ling, Z., Liu, L., Jiang, Y., Wu, H., Dai, L.: End-to-end emotional speech synthesis using style tokens and semisupervised training. In: Asia-Pacific Signal and Information Processing Association Annual Samit and Conf. (APSIPA ASC), pp. 623–627. IEEE (2019)
42. Zhou, K., Sisman, B., Rana, R., Schuller, B.W., Li, H.: Speech synthesis with mixed emotions. IEEE Trans. Affect. Comput. **14**(4), 3120–3134 (2022)
43. Van den Oord, A., Dieleman, S., Zen, H., et al.: WaveNet: a generative model for raw audio. arXiv preprint arXiv:1609.03499 12 (2016)
44. Van den Oord, A., et al.: Parallel WaveNet: fast high- fidelity speech synthesis. In: Proceedings of the 35th International Conference on Machine Learning, pp. 3915–3923. Stockholm, Sweden (2018)
45. Arik, S.O., et al.: Deep voice: real-time neural text-to-speech. In: Proceedings of the 34th International Conference on Machine Learning, pp. 195–204. PMLR, Sydney, Australia (2017)
46. Wang, Y., et al.: Tacotron: towards end-to-end speech synthesis. In: Proceedings of the 18th INTERSPEECH 2017, pp. 4006–4010. ISCA, Stockholm, Sweden (2017)
47. Shen, J., et al.: Natural TTS synthesis by conditioning WaveNet on MEL spectrogram predictions. In: Proceedings of ICASSP, pp. 4779–4783. Calgary, Canada (2018)
48. Ping, W., Peng, K., Gibiansky, A., et al.: Deep voice 3: scaling text-to-speech with convolutional sequence learning. arXiv preprint arXiv:1710.07654 (2017)
49. Arik, S.Ö, Chen, J., Peng, K., Ping, W., Zhou, Y.: Neural voice cloning with a few samples. In: Advances in Neural Information Processing Systems 31, 32nd Conference on Neural Information Processing Systems, pp. 10040–10050, Montreal, Canada (2018)
50. Nachmani, E., Polyak, A., Taigman, Y., Wolf, L.: Fitting new speakers based on a short untranscribed sample. In: Proceedings of the 35th International Conference on Machine Learning, pp. 3680–3688. Stockholm, Sweden (2018)
51. Akuzawa, K., Iwasawa, Y., Matsuo, Y.: Expressive speech synthesis via modeling expressions with variational autoencoder. In: Proceedings of the 19th INTERSPEECH, pp. 3067–3071. ISCA, Hyderabad, India (2018)
52. Ren, Y., et al.: Fastspeech: fast, robust and controllable text to speech. Adv. Neural Inf. Process. systems **32** (2019)
53. Ren, Y., et al.: Fastspeech 2: Fast and high-quality end-to-end text to speech. Preprint arXiv: 2006.04558 (2020)
54. Nosek, T., Suzić, S., Sečujski, M., Stanojev, V., Pekar, D., Delić, V.: End-to-end speech synthesis for the Serbian language based on Tacotron. In: Karpov, A. Delić, V., (eds.) SPECOM 2024, LNAI Part I - 15299, Springer, Heidelberg, Belgrade, Serbia (2024)
55. Wang, C., et al.: Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. arXiv preprint arXiv:2301.02111 (2023)
56. Zhang, Z., et al.: Speak foreign languages with your own voice: Cross-lingual neural codec language modeling. arXiv preprint arXiv:2303.03926 (2023)

57. Han, B., et al.: VALL-E R: Robust and Efficient Zero-Shot Text-to-Speech Synthesis via Monotonic Alignment. arXiv preprint arXiv:2406.07855 (2024)
58. Meng, L., et al.: Autoregressive Speech Synthesis without Vector Quantization. arXiv preprint arXiv:2407.08551 (2024)
59. Casanova, E., Weber, J., Shulby, C., Candido Junior, A., Gölge, E., Antonelli Ponti, M.: YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for everyone. arXiv preprint arXiv:2112.02418 (2024)
60. Kong, J., Kim, J., Bae, J.: HiFi-GAN: generative adversarial networks for efficient and high fidelity speech synthesis. arXiv preprint arXiv:2010.05646 (2020)
61. Prenger, R., Valle, R., Catanzaro, B.: WaveGlow: A Flow-based Generative Network for Speech Synthesis. arXiv preprint arXiv:1811.00002 (2018)
62. Casanova, E., et al.: XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model. arXiv preprint arXiv:2406.04904 (2024)
63. Sečujski, M., Obradović, R., Pekar, D., Jovanov, Lj., Delić, V.: AlfaNum system for speech synthesis in Serbian language. In: Proceedings of the 5th International Conference Text, Speech and Dialogue (TSD 2002), pp. 237–244. Brno, Czech Republic (2002)
64. Pakoci, E., Mak, R.: HMM-based speech synthesis for the Serbian language. In: Proceedings of the 56th ETRAN, vol. TE4, pp. 1–4. Zlatibor, Serbia (2012)
65. Delić, T., Sečujski, M., Suzić, S.: A review of serbian parametric speech synthesis based on deep neural networks. TELFOR J. **9**(1), 32–37 (2017)
66. Sečujski, M., Pekar, D., Suzić, S., Smirnov, A., Nosek, T.: Speaker/style-dependent neural network speech synthesis based on speaker/style embedding. J. Univ. Comput. Sci. **26**(4), 434–453 (2020)
67. Suzić, S., Sečujski, M., Nosek, T., Delić, V., Pekar, D.: HiFi-GAN based text-to-speech synthesis in Serbian. In: Proceedings of 30th EUSIPCO, pp. 2231–2235, Belgrade, Serbia (2022)
68. Sakai, T., Doshita, S.: Phonetic Typewriter. J. Acoust. Soc. Am. **33**, 1664 (1961)
69. Davis, K.H., Biddulph, R., Balashek, S.: Automatic recognition of spoken digits. J. Acoust. Soc. Am. **24**, 637–642 (1952)
70. Vintsyuk, T.K.: Speech discrimination by dynamic programming. Cybern. Syst. Anal. **4**, 52–57 (1972)
71. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans. Acoust. Speech Signal Process. **26**, 43–49 (1978)
72. Atal, B.S., Hanauer, S.L.: Speech analysis and synthesis by linear prediction of the speech wave. J. Acoust. Soc. Am. **50**, 637–655 (1971)
73. Jelinek, F., Bahl, L., Mercer, R.: Design of a linguistic statistical decoder for the recognition of continuous speech. IEEE Trans. Inf. Theory **21**, 250–256 (1975)
74. Klatt, D.H.: Review of the ARPA speech understanding project. J. Acoust. Soc. Am. **62**, 1345–1366 (1977)
75. Jelinek, F.: Continuous speech recognition by statistical methods. Proc. IEEE **64**, 532–556 (1976)
76. Levinson, S.E., Rabiner, L.R., Sondhi, M.M.: An Introduction to the application of the theory of probabilistic functions of a markov process to automatic speech recognition. Bell Syst. Tech. J. **62**, 1035–1074 (1983)
77. Juang, B.-H.: Maximum-likelihood estimation for mixture multivariate stochastic observations of markov chains. AT&T Tech. J. **64**, 1235–1249 (1985)
78. Juang, B.-H., Levinson, S., Sondhi, M.: Maximum likelihood estimation for multivariate mixture observations of Markov chains. IEEE Trans. on Inform. Theory **32**, 307–309 (1986)
79. Lee, K.-F.: Context-independent phonetic hidden Markov models for speaker-independent continuous speech recognition. IEEE Trans. Acoust. Speech Signal Process. **38**, 599–609 (1990)

80. Young, S.J., Woodland, P.C.: State clustering in hidden Markov model-based continuous speech recognition. Comput. Speech Lang. **8**, 369–383 (1994)
81. Mermelstein, P.: Distance measures for speech recognition, psychological and instrumental. Pattern Recogn. Artif. Intell. 374–388 (1976)
82. Hermansky, H.: Perceptual linear predictive (PLP) analysis of speech. J. Acoust. Soc. Am. **87**, 1738–1752 (1990)
83. Viikki, O., Laurila, K.: Cepstral domain segmental feature vector normalization for noise robust speech recognition. Speech Commun. **25**, 133–147 (1998)
84. Prasad, N.V., Umesh, S.: Improved cepstral mean and variance normalization using Bayesian framework. In: IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 156–161. IEEE, Olomouc, Czech Republic (2013)
85. Rehr, R., Gerkmann, T.: Cepstral noise subtraction for robust automatic speech recognition. In: Proceedings of ICASSP, pp. 375–378. IEEE, South Brisbane, Queensland, Australia (2015)
86. Hermansky, H., Morgan, N.: RASTA processing of speech. IEEE Trans. on Speech Audio Processing **2**, 578–589 (1994)
87. Bahl, L., Brown, P., De Souza, P., Mercer, R.: Maximum mutual information estimation of hidden Markov model parameters for speech recognition. In: Proceedings of ICASSP, pp. 49–52. IEEE, Tokyo, Japan (1986)
88. Valtchev, V., Odell, J.J., Woodland, P.C., Young, S.J.: MMIE training of large vocabulary recognition systems. Speech Commun. **22**, 303–314 (1997)
89. Juang, B.-H., Hou, W., Lee, C.-H.: Minimum classification error rate methods for speech recognition. IEEE Trans. Speech Audio Process. **5**, 257–265 (1997)
90. Povey, D., Woodland, P.C.: Minimum phone error and i-smoothing for improved discriminative training. In: Proceedings of ICASSP, pp. I-105-I–108. IEEE, Orlando, FL, USA (2002)
91. Ng, A.Y., Jordan, M.I.: On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes. In: Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic, pp. 841–848. MIT Press, Cambridge, MA, USA (2001)
92. Macherey, W.: Discriminative training and acoustic modeling for automatic speech recognition. Ph.D. Thesis, Aachen Techn. Hochsch (2010)
93. Baker, J.: The DRAGON system–An overview. IEEE Trans. Acoust. Speech Signal Process. **23**, 24–29 (1975)
94. Bahl, L.R., Jelinek, F., Mercer, R.L.: A maximum likelihood approach to continuous speech recognition. IEEE Trans. Pattern Anal. Mach. Intell. PAMI-5, 179–190 (1983)
95. Chen, S.F., Goodman, J.: An empirical study of smoothing techniques for language modeling. Comput. Speech Lang. **13**, 359–393 (1999)
96. Goodman, J.T.: A bit of progress in language modeling. Comput. Speech Lang. **15**, 403–434 (2001)
97. Lippmann, R.P.: Review of neural networks for speech recognition. Neural Comput. **1**, 1–38 (1989)
98. Bourlard, H.A., Morgan, N.: Connectionist Speech Recognition: a Hybrid Approach. Springer, US, Boston, MA (1994)
99. Mohamed, A., Dahl, G.E., Hinton, G.E.: Deep belief networks for phone recognition. In: NIPS Workshop on Deep Learning for Speech Recognition and Related Applications, pp. 1–9. Vancouver, BC, Canada (2009)
100. Dahl, G.E., Dong Yu, Li Deng, Acero, A.: Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans. Audio, Speech, Lang. Process. **20**, 30–42 (2012)

101. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the International Conference on Machine Learning, pp. 369–376. ACM Press, Pittsburgh, Pennsylvania (2006)
102. Maas, A., Xie, Z., Jurafsky, D., Ng, A.: Lexicon-free conversational speech recognition with neural networks. In: Proceedings Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 345–354. Denver, Colorado (2015)
103. Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P., Bengio, Y.: End-to-end attention-based large vocabulary speech recognition. In: Proceedings of ICASSP, pp. 4945–4949. Shanghai (2016)
104. Karita, S., et al.: A comparative study on transformer vs RNN in speech applications. In: Automatic speech recognition and understanding workshop (ASRU), pp. 449–456. IEEE, SG, Singapore (2019)
105. Zhu, H., Wang, L., Cheng, G., Wang, J., Zhang, P., Yan, Y.: Wav2vec-S: semi-supervised pre-training for low-resource ASR. In: Proceedings of the 23th INTERSPEECH, pp. 4870–4874. ISCA (2022)
106. Schneider, S., Baevski, A., Collobert, R., Auli, M.: wav2vec: Unsupervised pre-training for speech recognition. arXiv preprint arXiv:1904.05862 (2019)
107. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: Proceedings of the International Conference on Machine Learning, pp. 28492–28518 (2023)
108. Suzić, S., Ostrogonac, S., Pakoci, E., Bojanić, M.: Building a speech repository for a Serbian LVCSR system. Telfor J. **6**(2), 109–114 (2014)
109. Nosek, T., Suzić, S., Delić, V., Sečujski, M.: Cross-lingual text-to-speech with prosody embedding. In: Proceedings of IWSSIP, 5 pages (2023)
110. Pakoci, E.T., Popović, B.Z.: Recurrent neural networks and morphological features in language modeling for Serbian. In: 29th Telecommunication Forum (TELFOR), 8 pages. IEEE (2021)
111. Delić, V., Sečujski, M., Sedlar, N.V., Mišković, D., Mak, R., Bojanić, M.: How speech technologies can help people with disabilities. In: Ronzhin, A., Potapova, R., Delić, V. (eds.) 16th SPECOM 2014, LNAI, vol. 8773, pp. 243–250. Springer. Novi Sad, Serbia (2014)
112. Delić, V., et al.: Central audio-library of the university of Novi Sad. In: Proceedings of the Intelligent Distributed Computing XIII, pp. 467–476. Springer International Publishing (2020)
113. Pakoci, E., Pekar, D., Popović, B., Sečujski, M., Delić, V.: Overcoming data sparsity in automatic transcription of dictated medical findings. In: Proceedings of the 30th EUSIPCO, pp. 454–458. IEEE (2022)
114. Popović, B., Pakoci, E., Jakovljević, N., Kočiš, G., Pekar, D.: Voice assistant application for the Serbian language. In: 23rd Telecommunication Forum (TELFOR), pp. 858–861. IEEE (2015)
115. Reitmaier, T., et al: Opportunities and challenges of automatic speech recognition systems for low-resource language speakers. In Proceedings of the CHI Conference on Human Factors in Computing Systems, p. 17 (2022)
116. Mu, Z., Yang, X., Dong, Y.: Review of end-to-end speech synthesis technology based on deep learning. arXiv preprint arXiv:2104.09995 (2021)
117. Ogayo, P., Neubig, G., Black, A.W.: Building TTS systems for low resource languages under resource constraints. In: Proceedings Speech for Social Good Workshop, p. 5 (2022)
118. Jimerson, R., Liu, Z., Prud'Hommeaux, E.: An (unhelpful) guide to selecting the best ASR architecture for your under-resourced language. In: Proceedings of the 61st Annual Meeting of the Association for Comp. Linguistics (Vol. 2 Short Papers), pp. 1008–1016 (2023)

119. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: Wav2vec 2.0: a framework for self-supervised learning of speech representations. Adv. Neural Inf. Process. Syst. **33**, 12449–12460 (2020)
120. Popović, B.Z., Pakoci, E.T., Pekar, D.J.: Transfer learning for domain and environment adaptation in Serbian ASR. Telfor Journal **12**(2), 110–115 (2020)
121. Delić, V.D., Pekar, D.J., Sečujski, M.S., Popović, B.Z., Pakoci, E.T., Suzić, S.B.: Development of speech technology for Serbian and its applications. In: Proceedings of the First Serbian International Conference on Applied Artificial Intelligence, p. 7. Kragujevac, Serbia (2022)

Automatic Speech Recognition

Comparison of Well and Lower-Resourced Self-training in ASR

Yue Luo[1,2(✉)] and Péter Mihajlik[1,3]

[1] Departure of Telecommunications and Artificial Intelligence, Budapest University of Technology and Economics, Budapest, Hungary
luo.yue@edu.bme.hu, mihajlik@tmit.bme.hu
[2] SpeechTex Ltd., Budapest, Hungary
[3] HUN-REN Hungarian Research Centre for Linguistics, Budapest, Hungary

Abstract. In this paper, we present a case study of self-training for end-to-end Hungarian and Mandarin speech recognition. We demonstrate that self-training methods can significantly improve the accuracy of the baseline model for both languages. We trained a supervised baseline model which achieved a 20.13% WER on the BEA-Base eval-spont set. The self-trained Hungarian model, combining pseudo-labels generated by the baseline seed model with labeled data, achieved a WER of 15.93% on the BEA-Base eval-spont set, representing a 20.84% relative reduction compared to the baseline, and reduced WER by relative 15.28% on the independent Common Voice test set. The Mandarin model, which relied solely on pseudo-labels from the Whisper large-V2 model and used no labeled data, reduced CER by 45.31% on the AISHELL-2018A-EVAL test set, improving from 13.00% to 7.11%, and by relative 32.42% on the external Common Voice test set. We find that for spontaneous, low-resource Hungarian ASR tasks, pseudo-labels from domain-specific models are more effective than those from large general models like Whisper large-V2.

Keywords: Automatic speech recognition · Self-training · Semi-supervised training · Hungarian · Mandarin

1 Introduction

Automatic Speech Recognition (ASR) has advanced significantly in recent years, particularly with the development of end-to-end ASR frameworks [4,10]. These frameworks have simplified the traditional ASR pipeline into a single neural network [26], achieving notable success by directly converting audio signals into text.

However, a major challenge is that end-to-end ASR systems rely heavily on large amounts of labeled data [16]. Obtaining labeled data is expensive and time-consuming [15], and the difficulty is particularly pronounced for low-resource languages like Hungarian, which have limited labeled data available [23]. Therefore,

A. Karpov and V. Delić (Eds.): SPECOM 2024, LNAI 15299, pp. 45–56, 2025.
https://doi.org/10.1007/978-3-031-77961-9_3

semi-supervised learning becomes a promising approach for end-to-end ASR systems [30] as it can leverage the abundance of unpaired audio and text data [31]. Among the various semi-supervised learning methods [38], self-training [6,28](or pseudo-labeling [19]) stands out due to its simplicity and effectiveness [13,21].

Self-training involves using a seed model to generate pseudo-labels (automatic transcriptions) for unlabeled data, thereby enlarging the training set for model training [13]. A seed model can be trained from scratch using a limited amount of labeled data or by utilizing the pre-trained models directly. With the development of multilingual speech recognition models like Whisper [27] and Universal Speech Model (USM) [40], these pre-trained models can also serve as seed models.

Our study aims to evaluate the effectiveness of self-training methods by assessing how self-training improves the performance of ASR systems with limited labeled data, particularly for Hungarian and Mandarin. To achieve the objective, we conducted several experiments involving recent neural architectures and training schemes, such as Conformer [11] and Transformer [32]. Our experiments included supervised training from scratch, as well as self-training strategies that combine pseudo-labels with labeled data and those that exclusively use pseudo-labels. The models were evaluated on test sets for Hungarian and Mandarin to determine the effectiveness of self-training and its performance across different resource conditions. Additionally, for a spontaneous, lower-resourced Hungarian ASR task, we compared the impact of pseudo-labels generated by different seed models (trained with a small labeled dataset versus pre-trained models like faster Whisper large-V2[1]) on model performance, assessing the importance of seed model selection and pseudo-label quality on ASR system performance.

The paper is organized as follows: Section 2 reviews the progress of semi-supervised learning techniques in ASR, focusing on self-training methods. Section 3 describes the self-training method used in our experiments, along with the different strategies employed for self-training in Hungarian and Mandarin. Section 4 presents the datasets and experimental setup. Section 5 provides the results of baseline models and models applying different self-training methods to Hungarian and Mandarin and discusses the results of the study, followed by concluding remarks in the last section.

2 Related Work

The early work of Shahshahani and Landgrebe [29] recognized the value of unlabeled data and is often regarded as the starting point for semi-supervised learning research. Notable applications in ASR began in 2004 when Kamm et al. [18] used semi-supervised learning for acoustic model training. As the most typically used approach in semi-supervised learning [28], self-training has also been intensively studied by scholars in the field of ASR.

[1] https://huggingface.co/Systran/faster-whisper-large-v2.

A Study [33] demonstrated that iteratively regenerating pseudo-labels and retraining models, while comparing two data filtering methods, can enhance model performance. Kahn et al. [15] introduced a self-training approach for end-to-end ASR models that uses pseudo-label filtering and language model decoding to improve word error rates. In 2019, [25] introduced curriculum learning into semi-supervised learning, allowing models to alternate learning from labeled and unlabeled data. Park et al. [24] adopted the Noisy Student Training (NST) [36] method, integrating it with the adaptive SpecAugment data augmentation technique. The superiority of combining self-training with pre-training was validated in [8], achieving significant performance boosts. Additionally, SentAugment [8] was proposed to reduce noise from general corpora. In [37], Xu et al. combined self-training and pre-training, showing that pre-trained self-supervised methods like wav2vec are complementary to self-training. A strategy introduced by [14] balances pseudo-label quality and quantity by using average probability scores from model outputs to filter low-quality pseudo-labels.

Compared with well-resource languages such as English and Mandarin, there are fewer research cases on semi-supervised learning for Hungarian ASR. [20] leverages CycleGAN and inter-domain loss to improve noisy student training, thereby improving the quality of probabilistic transcripts generated from a limited Hungarian data source. In addition, the Hungarian ASR model that implements a semi-supervised learning strategy is mentioned as a baseline model in [5,12]. Although these studies employ various semi-supervised methods to enhance Hungarian ASR performance, the baseline models used in previous research were not strong. In contrast, our baseline model is more competitive, and our approach significantly outperforms the strong baseline.

3 Methods

3.1 Supervised Models

In the supervised learning approach, we leverage fully labeled datasets to train our ASR models. This approach serves as a baseline for comparing the effectiveness of semi-supervised strategies. Additionally, these supervised models can act as seed models in subsequent self-training phases.

For the Hungarian supervised model, we used the BEA-Base dataset [22]. For Mandarin, we developed two supervised models using the AISHELL-1 [3] and AISHELL-2 [7] datasets. The WeNet [39] toolkit was employed to train these models. Once trained, the models were evaluated on the Common Voice-17.0-hu [1] test set and BEA-Base eval-spont set for Hungarian, and on the Common Voice-17.0-zh [1] test set and AISHELL-2018A-EVAL test set for Mandarin. Further experimental setup details will be provided in Sect. 4.

3.2 Semi-supervised Models

In the experiment, we employed various self-training strategies to develop semi-supervised ASR models for Hungarian and Mandarin. The processes involve two

main phases (if a pre-trained model is not used as the seed model): the initial training phase and the iterative training phase, as illustrated in Fig. 1. During the iterative training phase, the key difference between the two strategies in the experiment lies in the use of data: one strategy combines pseudo-labels with labeled data (PLL), while the other strategy only uses pseudo-labels (PL).

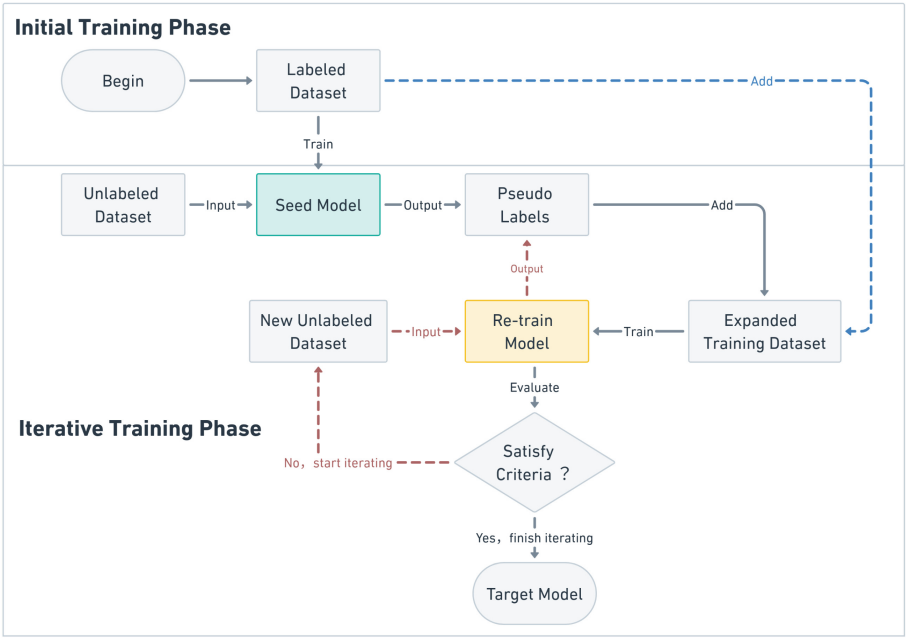


Fig. 1. The illustration of self-training method in ASR.

During the "Initial Training Phase", we start by creating a seed model using a labeled dataset $D_{\text{labeled}} = \{(X_l, Y_l)\}$. This involves training an initial ASR model C on the Hungarian BEA-Base dataset. This model serves as the foundation for generating pseudo-labels in subsequent stages. The trained seed model C is then used to transcribe a large set of unlabeled audio data $D_{\text{unlabeled}} = \{X_u\}$, resulting in a set of pseudo-labels $\hat{Y}_u$. Despite being generated by a model with limited initial accuracy, these pseudo-labels allow us to expand our training dataset without requiring additional labeled data.

Pseudo-Labels and Labeled Data (PLL): In the "Iterative Training Phase", the pseudo-labeled data $D'_{\text{unlabeled}} = \{(X_u, \hat{Y}_u)\}$ is combined with the original labeled dataset to form an expanded training set $D_{\text{expanded}} = D_{\text{labeled}} \cup D'_{\text{unlabeled}}$. This new training dataset, which includes both genuine and pseudo-labels, is used to train a new ASR model C_{new}.

In our experiment, this process was executed a single time. Nonetheless, the methodology supports iterative execution. In each subsequent iteration i, the ASR model C_{i-1} would transcribe additional unlabeled data, yielding improved pseudo-labels $\hat{Y}_u^{(i)}$ due to the enhanced performance of the model. These improved pseudo-labels would be combined with the labeled dataset to refine the expanded training set $D_{\text{expanded}}^{(i)} = D_{\text{labeled}} \cup D_{\text{unlabeled}}^{\prime(i)}$. The new ASR model C_i would then be trained using $D_{\text{expanded}}^{(i)}$.

Each iteration involves evaluating the model against specific performance criteria. If the model meets these criteria, the process conclude, resulting in $C_{\text{final}} = C_i$. If not, the iterative training would continue, benefiting from the increasingly accurate pseudo-labels generated by the previous models. Ultimately, this would produce an ASR model C_{final} which has undergone multiple rounds of self-training.

Pseudo-Labels (PL): In the "Iterative Training Phase" of the strategy that exclusively uses pseudo-labels, the process focuses on utilizing pseudo-labeled data alone to train subsequent ASR models. The seed model C initially generates pseudo-labels $\hat{Y}_u$ for the unlabeled dataset $D_{\text{unlabeled}} = \{X_u\}$, creating a pseudo-labeled dataset $D_{\text{pseudo}} = \{(X_u, \hat{Y}_u)\}$.

In our study, we applied this process only once. However, the approach is designed to be iterative. In each potential iteration i, the ASR model C_{i-1} would be trained solely on the pseudo-labeled dataset $D_{\text{pseudo}}^{(i-1)}$, which consists of audio inputs and their corresponding pseudo-labels generated in the previous iteration. This model C_i would then be used to transcribe additional unlabeled data, producing an improved set of pseudo-labels $\hat{Y}_u^{(i)}$. The newly pseudo-labeled dataset $D_{\text{pseudo}}^{(i)} = \{(X_u, \hat{Y}_u^{(i)})\}$ would replace the previous iteration's pseudo-labeled data.

This iterative cycle could be continued, with each new ASR model C_i improving its transcription accuracy using the refined pseudo-labels from the previous iteration. Performance evaluation criteria would guide the process, if the model's performance meets these criteria, the training process would conclude, resulting in the final ASR model $C_{\text{final}} = C_i$. If not, the iterations would persist, progressively improving the model's accuracy through better pseudo-label generation. Thus, the final ASR model C_{final} would be developed through exclusive reliance on pseudo-labels.

4 Experiment

4.1 Datasets

To train our model, we utilized a variety of labeled and unlabeled datasets in both Hungarian and Mandarin. For Hungarian, the labeled BEA-Base dataset provided a comprehensive corpus for training and evaluation, while the unlabeled BEA-Wavs dataset was used to generate pseudo-labels for the semi-supervised

learning process. Additionally, the Common Voice-17.0-hu dataset was employed to test the Hungarian model's performance.

For Mandarin, the AISHELL-1 and AISHELL-2 datasets, which are well-established labeled datasets containing extensive Mandarin speech data, were used in the training process. The AISHELL-2018A-EVAL dataset and Common Voice-17.0-zh dataset were utilized during the evaluation phase to test the recognition performance. These datasets collectively ensured robust training and evaluation, enhancing the overall efficacy of models.

BEA-Base[2] is a subset of the BEA dataset optimized for ASR. In order to intentionally exclude repeated and read sentences from the training set, the development set and evaluation set divide the data into spontaneous (naturally occurring or unprepared spoken expressions) and repeated (prepared readings or practiced utterances) categories, only "dev-spont" and "eval-spont (BEA-eval)" are used in the experiment.

BEA-Wavs derived from the BEA [9] ("BEszélt nyelvi Adatbázis" in Hungarian, meaning spoken language database) with recordings of 470 speakers uses Voice Activity Detection (VAD) to segment long audio files into about 30-second clips without transcripts.

AISHELL-1[3] consists of 178 h of quiet door environment recoded data, it provides a vast coverage for different Chinese accent regions with 400 speakers across 11 domains.

AISHELL-2[4] contains 1000 h of high-quality audio from 1991 speakers. In this study, it is also treated as unlabeled dataset by removing transcripts and then used to generate pseudo-labels for the semi-supervised training process.

AISHELL-2018A-EVAL[5] is a dataset of 5000 utterances from 10 speakers and development data with 2500 utterances from 5 speakers.

The Common Voice[6] created through crowd contributions and publicly accessible features speech samples in numerous languages. For the experiment's testing phase, the Chinese and Hungarian test sets from this dataset were employed (Table 1).

[2] https://phon.nytud.hu/bea/bea-base.html?lang=hu.
[3] https://www.aishelltech.com/kysjcp.
[4] https://www.aishelltech.com/aishell_2.
[5] https://www.aishelltech.com/aishell_2018_eval.
[6] https://commonvoice.mozilla.org/en/datasets.

4.2 Experimental Setups

In this experiment, the WeNet toolkit v2.2.0 was utilized in all setups, the machine configuration included an AMD Ryzen 9 5900X 12-Core Processor, 64GB memory, NVIDIA RTX A6000 GPU, and Ubuntu 20.04.6.

Table 1. The statistics of dataset used in experiment.

Abbrv.	Dataset	Train Hours	Dev Hours	Test Hours	Total Speaker	Used Subsets
BEA-Base	BEA-Base	71.2	4.02	4.91	140	train, dev, test
BEA-Wavs	BEA-Wavs	373.2	–	–	500	train
CV-hu	Common Voice-17.0-hu	53.35	16.68	17.73	1614	test
AI-1	AISELL-1	150	18	10	400	train, dev
AI-2	AISELL-2	1000	–	–	1991	train
AI-eval	AISELL-2018A-EVAL	–	2.03	3.54	15	dev, test
CV-zh	Common Voice-17.0-zh	42.34	15.92	17.45	3333	test

Initially, the dataset was downloaded, extracted, and prepared by organizing audio and text files, followed by computing Cepstral Mean and Variance Normalization (CMVN) [34] statistics. A dictionary was then created to map text tokens to indices, which was essential for model training. The training configuration specified a batch size of 4, with gradient accumulation set to 4 and the model was trained for 150 to 240 epochs. Feature extraction involved computing 80 Mel-frequency cepstral coefficients (MFCCs), and data augmentation techniques such as speed perturbation and spectral augmentation were applied to enhance the training data.

The model architecture consisted of a Conformer encoder with 12 blocks and a Transformer decoder with 6 blocks. The encoder had an output size of 256, four attention heads, 2048 linear units, and a dropout rates of 0.1 for both general and positional layers, a convolutional module with a kernel size of 15, swish activation function, and relative positional encoding. Similarly, the decoder was configured with four attention heads, 2048 linear units, 6 blocks, and dropout rates of 0.1.

The hybrid CTC/attention model [35] configuration included a CTC weight of 0.3 and a label smoothing weight of 0.1. For dataset processing, various configurations were applied, including filtering with specified max and min lengths, resampling at 16000 Hz, applying speed perturbation, extracting 80-dimensional MFCCs, using spectral augmentation with two time masks and two frequency masks, shuffling and sorting with specified sizes, and static batching with a batch size of 4.

Optimization was carried out using the Adam [17] optimizer with a learning rate of 0.002 and a warm-up learning rate scheduler with 25,000 warm-up steps, while gradient clipping was set to 5. After training, the model was evaluated

using an attention-based decoding strategy. Performance was measured using Word Error Rate (WER) to assess the accuracy of the Hungarian model and Character Error Rate (CER) for Mandarin.

5 Results and Discussion

We report the results in Tables 2 and 3. In experiments, we trained supervised models S_L_BEA-Base, S_L_AI-1, and S_L_AI-2 on labeled datasets for Hungarian and Mandarin, respectively. The S_L_BEA-Base model achieves a WER of 20.13% on BEA-Base eval-spont, which is a significant improvement compared to the previous benchmark [22], highlighting the effectiveness of labeled data. In addition, in the model that implements the self-training method, the ST_PLL_BEA-Base is trained on a combination of BEA-Wavs pseudo-labels, which are generated by the seed model S_L_BEA-Base, and BEA-Base label data. It achieves a WER of 15.93% on BEA-Base eval-spont, which is comparable to the performance of the "Mega" model [2,22]. However, the performance of both the ST_PLL_Whisper model, which is trained using a combination of BEA-Wavs pseudo-labels generated by the Whisper large V2 model and BEA-Base labeled data, and the ST_PL_BEA-Base model, which is trained only using the BEA-Wavs pseudo-labels generated by the S_L_BEA-Base model, has dropped significantly on different test sets.

Table 2. WERs of Hungarian Models with Different ML Methods and Training Dataset Configurations. Supervised learning (S) and self-training (ST) methods are implemented on different models. Different seed models, S_L_BEA-Base and Whisper large-V2 [27] are used to transcribe the unlabeled dataset BEA-Wavs to generate pseudo-labels. These pseudo-labels are either used directly as training sets or combined with the labeled dataset BEA-Base to form training sets.

Model	Training data size [hours]	**Hungarian (WER[%])**	
		BEA-eval	CV-hu
S_L_BEA-Base	71.2	20.13	37.95
ST_PL_BEA-Base	373.2	31.93	63.04
ST_PLL_BEA-Base	444.4	**15.93**	**32.16**
ST_PLL_Whisper	444.4	31.35	56.69

Notably, unlike the Hungarian results, Mandarin has a similar CER for models trained under both self-training strategies (pseudo-labels only and pseudo-labels combined with labels). Compared with the baseline model trained with a small amount of labeled data (AISHELL-1), using the self-training method significantly improves the performance of the model. On the AISHELL-2018A-EVAL test set, the CER is reduced by 45.31%, from 13.00% to 7.11%. Similarly, on the Common Voice test set, the CER is reduced by 32.42%, from 35.87%

Table 3. CERs of Mandarin Models with Different ML Methods and Training Dataset Configurations. Supervised learning (S) and self-training (ST) methods are implemented on different models. Whisper large-V2 is used to transcribe the AISHELL-2 audio files to generate pseudo-labels. These pseudo-labels are either directly used as training sets or combined with the labeled dataset AISHELL-1 to form training sets.

Model	Training data size [hours]	**Mandarin (CER[%])**	
		AI-eval	CV-zh
S_L_AI-1	150	13.00	35.87
S_L_AI-2	1000	6.36	27.15
ST_PL_Whisper	1000	**7.03**	**24.51**
ST_PLL_Whisper	1150	7.11	24.24

to 24.24%. Furthermore, the model trained using only pseudo-labels (unlabeled data) generated by Whisper large-V2 is equivalent to the model trained using a large amount of labeled data (AISHELL2) on the AISHELL-2018A-EVAL test set. On the Common Voice test set, the former effect is even better than the latter.

This difference between Hungarian and Mandarin highlights several key aspects. In Hungarian, among the models implementing self-training, only the ST_PLL_BEA-Base model achieved a reduction in WER compared to the baseline, indicating that self-training can be effective with limited labeled data, highlighting the importance of combining pseudo-labeled data with label data to correct errors and improve performance. However, the choice of seed model and the quality of pseudo-labeled data play a crucial role. The lower performance of the ST_PLL_Whisper model on BEA-Base compared to the ST_PLL_BEA-Base model is likely due to the mismatch between the Whisper model's training data (mainly transcription and translation data) and the spontaneous nature of BEA-Base dataset. This makes it difficult for Whisper models trained on out-of-domain data to generate accurate pseudo-labels.

The superior performance of the Mandarin ASR system trained using only pseudo-labels can be attributed to the diversity and quality of the pseudo-label data generated by Whisper large-V2. After extensive training on various transcription and translation tasks, the Whisper model provides rich and powerful representations that are beneficial for Mandarin ASR. Additionally, the structured nature of the AISHELL dataset (consisting of spoken speech in a quiet environment) may contribute to the effectiveness of pseudo-labeling, as Whisper is well-suited for such data.

Moreover, for both Hungarian and Mandarin, the variation in WER(CER) between the respective homologous test sets and the external Common Voice datasets highlights the challenges in generalization across different datasets. This discrepancy underscores the importance of dataset diversity in training robust speech recognition models.

6 Conclusion

In this paper, we explored how different self-training strategies improve ASR performance, with clear differences between Hungarian and Mandarin. For Hungarian, self-training showed clear benefits in improving accuracy, but its success depended heavily on the quality of pseudo-labels and the choice of the seed model. The spontaneous nature of the Hungarian BEA-Base dataset posed significant challenges, underscoring the importance of domain-specific tuning and the relevance of pseudo-labeled data to the target domain.

In contrast, Mandarin demonstrated significant performance improvements by applying self-training methods, largely due to the high quality and diversity of pseudo-labels generated by the Whisper large-V2 model and the structured nature of the AISHELL dataset. These factors facilitated more effective use of pseudo-labels, leading to substantial reductions in CER across different test sets.

The study also reveals that generalizing across different datasets remains a challenge, as shown by variations in WER(CER) between homologous test sets and external datasets. This underscores the importance of dataset diversity in developing robust ASR systems.

Overall, self-training emerges as a powerful method for improving ASR performance in low-resource settings. However, its effectiveness is influenced by factors such as the quality of pseudo-labels, the choice of the seed model, and the relevance of the data to the target domain. These findings provide valuable insights into optimizing self-training approaches for different languages and speech characteristics, contributing to the advancement of ASR technologies in diverse linguistic contexts.

Acknowledgments. The research was supported partially by the NKFIH K143075, K135038 and NKFIH-828-2/2021(MILAB) projects of the NRDI Fund.

References

1. Ardila, R., et al.: Common voice: a massively-multilingual speech corpus. arXiv preprint arXiv:1912.06670 (2019)
2. Babu, A., et al.: XLS-R: self-supervised cross-lingual speech representation learning at scale. arXiv preprint arXiv:2111.09296 (2021)
3. Bu, H., Du, J., Na, X., Wu, B., Zheng, H.: AISHELL-1: an open-source mandarin speech corpus and a speech recognition baseline. In: 2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA), pp. 1–5. IEEE (2017)
4. Chan, W., Jaitly, N., Le, Q., Vinyals, O.: Listen, attend and spell: a neural network for large vocabulary conversational speech recognition. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4960–4964. IEEE (2016)
5. Chen, W., Hasegawa-Johnson, M., Chen, N.F.: Recognizing zero-resourced languages based on mismatched machine transcriptions. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5979–5983. IEEE (2018)

6. Chen, Y., Wang, W., Wang, C.: Semi-supervised ASR by end-to-end self-training. arXiv preprint arXiv:2001.09128 (2020)
7. Du, J., Na, X., Liu, X., Bu, H.: AISHELL-2: transforming mandarin ASR research into industrial scale. arXiv preprint arXiv:1808.10583 (2018)
8. Du, J., et al.: Self-training improves pre-training for natural language understanding. arXiv preprint arXiv:2010.02194 (2020)
9. Gósy, M.: Bea-a multifunctional Hungarian spoken language database. Phonetician **105**, 50–61 (2013)
10. Graves, A., Jaitly, N.: Towards end-to-end speech recognition with recurrent neural networks. In: International Conference on Machine Learning, pp. 1764–1772. PMLR (2014)
11. Gulati, A., et al.: Conformer: convolution-augmented transformer for speech recognition. arXiv preprint arXiv:2005.08100 (2020)
12. Hasegawa-Johnson, M.A., et al.: ASR for under-resourced languages from probabilistic transcription. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(1), 50–63 (2016)
13. Huang, Y., Wang, Y., Gong, Y.: Semi-supervised training in deep learning acoustic model. In: Interspeech, pp. 3848–3852 (2016)
14. Jin, Z., Zhong, D., Song, X., Liu, Z., Ye, N., Zeng, Q.: Filter and evolve: progressive pseudo label refining for semi-supervised automatic speech recognition. arXiv preprint arXiv:2210.16318 (2022)
15. Kahn, J., Lee, A., Hannun, A.: Self-training for end-to-end speech recognition. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7084–7088. IEEE (2020)
16. Kahn, J., et al.: Libri-Light: a benchmark for ASR with limited or no supervision. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7669–7673. IEEE (2020)
17. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
18. Lawrence, N., Jordan, M.: Semi-supervised learning via gaussian processes. In: Advances in Neural Information Processing Systems, vol. 17 (2004)
19. Lee, D.H., et al.: Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on Challenges in Representation Learning, ICML, vol. 3, p. 896. Atlanta (2013)
20. Li, C.Y., Vu, N.T.: Improving noisy student training for low-resource languages in end-to-end ASR using CycleGAN and inter-domain losses. In: Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages@ LREC-COLING 2024, pp. 133–142 (2024)
21. Liao, H., McDermott, E., Senior, A.: Large scale deep neural network acoustic modeling with semi-supervised training data for Youtube video transcription. In: 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 368–373. IEEE (2013)
22. Mihajlik, P., Balog, A., Gráczi, T.E., Kohári, A., Tarján, B., Mády, K.: Bea-base: a benchmark for asr of spontaneous Hungarian. arXiv preprint arXiv:2202.00601 (2022)
23. Mihajlik, P., et al.: Is spoken Hungarian low-resource?: a quantitative survey of Hungarian speech data sets. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp. 9382–9388 (2024)
24. Park, D.S., et al.: Improved noisy student training for automatic speech recognition. arXiv preprint arXiv:2005.09629 (2020)

25. Parthasarathi, S.H.K., Strom, N.: Lessons from building acoustic models with a million hours of speech. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6670–6674. IEEE (2019)
26. Prabhavalkar, R., Hori, T., Sainath, T.N., Schlüter, R., Watanabe, S.: End-to-end speech recognition: A survey. Speech Lang. Process. IEEE/ACM Trans. Audio **32**, 325–351 (2023)
27. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: International Conference on Machine Learning, pp. 28492–28518. PMLR (2023)
28. Scudder, H.: Probability of error of some adaptive pattern-recognition machines. IEEE Trans. Inf. Theory **11**(3), 363–371 (1965)
29. Shahshahani, B.M., Landgrebe, D.A.: The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon. IEEE Trans. Geosci. Remote Sens. **32**(5), 1087–1095 (1994)
30. Synnaeve, G., et al.: End-to-end ASR: from supervised to semi-supervised learning with modern architectures. arXiv preprint arXiv:1911.08460 (2019)
31. Van Engelen, J.E., Hoos, H.H.: A survey on semi-supervised learning. Mach. Learn. **109**(2), 373–440 (2020)
32. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
33. Veselỳ, K., Hannemann, M., Burget, L.: Semi-supervised training of deep neural networks. In: 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 267–272. IEEE (2013)
34. Viikki, O., Laurila, K.: Cepstral domain segmental feature vector normalization for noise robust speech recognition. Speech Commun. **25**(1–3), 133–147 (1998)
35. Watanabe, S., Hori, T., Kim, S., Hershey, J.R., Hayashi, T.: Hybrid CTC/attention architecture for end-to-end speech recognition. IEEE J. Sel. Top. Signal Process. **11**(8), 1240–1253 (2017)
36. Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: Self-training with noisy student improves ImageNet classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10687–10698 (2020)
37. Xu, Q., et al.: Self-training and pre-training are complementary for speech recognition. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3030–3034. IEEE (2021)
38. Yang, X., Song, Z., King, I., Xu, Z.: A survey on deep semi-supervised learning. IEEE Trans. Knowl. Data Eng. **35**(9), 8934–8954 (2022)
39. Yao, Z., et al.: WeNet: production oriented streaming and non-streaming end-to-end speech recognition toolkit. arXiv preprint arXiv:2102.01547 (2021)
40. Zhang, Y., et al.: Google USM: scaling automatic speech recognition beyond 100 languages. arXiv preprint arXiv:2303.01037 (2023)

Towards a Livvi-Karelian End-to-End ASR System

Irina Kipyatkova(✉), Ildar Kagirov, Mikhail Dolgushin, and Alexandra Rodionova

St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), 14th Line, 39, 199178 St. Petersburg, Russia
{kipyatkova,kagirov,dolgushin.m}@iias.spb.su, santrar@krc.karelia.ru

Abstract. This paper presents an investigation of the performance of various end-to-end ASR models trained on low-rescourced Livvi-Karelian. Several Wav2Vec 2.0 and Whisper based models were fine-tuned, tested and compared with the hybrid TDNN-F/HMM. In the course of the experiments, end-to-end Transformer-based models have demonstrated a good performance, however the best results obtained were due to a combination of N-gram and Transformer-based models. The result of 19.83% WER on the test set were obtained using the Wav2Vec 2.0 large model with N-gram augmentation, thus being on par with SOTA models for other low-resource languages. Besides, this paper presents a new language corpus of Livvi-Karelian, containing transcripts from radio broadcasts, featuring samples from 17 speakers (7 males and 10 females). Covering about 4.5 h of audio recordings, it contains 32,037 words, thus being a valuable tool for linguistic research. The findings of the presented work may be of considerable interest both for low-resource ASR and field Finnougristics.

Keywords: Livvi-Karelian · Speech Corpora · Code-Switching

1 Introduction

This paper presents the results of speech recognition experiments conducted on a new Livvi-Karelian (further also – Karelian) corpus AnKaS.[1] The Livvi-Karelian language is one of the main Karelian idioms spoken primarily in the Republic of Karelia (Russia) and some parts of Finland. It belongs to the Balto-Finnic branch of the Finno-Ugric group of the Uralic languages. Livvi-Karelian, also known as Olonets Karelian, differs from the Northern Karelian dialects in its phonology, vocabulary, and some grammatical structures, although it shares a close affinity with Finnish. The language has faced a decline in native speakers over the years, with 12,367 self-reported speakers in Russia (2010 census) and up to 20,000 in Finnland (according to the data as of the year

[1] AnKaS (Database of Annotations of Karelian Speech) can be found at https://irinakipyatkova.github.io/AnKaS/.

A. Karpov and V. Delić (Eds.): SPECOM 2024, LNAI 15299, pp. 57–68, 2025.
https://doi.org/10.1007/978-3-031-77961-9_4

2016) [1]. Livvi-Karelian (along with the other Karelian idioms) belongs to the so-called low-resource languages. A "low-resource language" is a language with severely limited electronic resources and data available for linguistic study and technological development [2, 3]. Currently, there is a great shortage of open-source databases of the Karelian speech for Automatic Speech Recognition (ASR) applications. The lack of speech data greatly hinders the development of state-of-the-art (SOTA) ASR systems for Karelian.

One of the ways to solve the problem of data shortage is to use a pre-trained model with a further fine-tuning on the target language speech data. Currently, there exist two SOTA pre-trained models which have gained popularity in the field of ASR, namely, the self-supervised Wav2Vec model and the supervised Whisper model.

Wav2Vec 2.0 was first introduced in [4]. The model was pre-trained and then fine-tuned on 1040 h of annotated English speech, and it surpassed previous SOTA results in speed, while requiring 11 times less annotated data. The authors of [5] analyzed the Wav2Vec system in the context of a multilingual pre-training approach. They showed that Cross-Lingual Speech Representations (XLSR) can be learnt and used by monolingual models. This idea was further developed in [6] using XLS-R models. In [7], the model was leveraged to more than 1000 languages, one of which was Karelian, due to effective usage of adapters and gradual addition of languages on the training stage [8]. Also, it was found out that the fuse of language models and linguistic encoders, such as BERT [9], with acoustic encoder might be promising for tasks of low-resource speech recognition, as it was shown in [10]. Some of these models, for example w2v-bert v2, can attain SOTA results [11] on ASR tasks.

ASR model Whisper developed by OpenAI [12] was trained using fully supervised methods, which involves using up to 680,000 h of labeled speech data from various sources. Due to the massive database and the training techniques used, the model can solely serve as a multilingual and multitask ASR system, sufficiently solving the tasks like language recognition, speech recognition and language translation in noisy environments. The model was enhanced to add the multitask training format using a set of special tokens that serve as task specifiers or classification targets.

The addition of an out-of-boundaries language or fine-tune of low-resourced language in Whisper might not be as effective as in the Wav2Vec-based models. For example, a comparison between Wav2Vec 2.0 and Whisper for low-resource Maltese ASR was presented in [13]. The authors performed experiments on fine-tuning Wav2Vec 2.0 XLS-R with 300 M and 2B, and Whisper-tiny, -small and -large models on Maltese-English code-switching data of different lengths (from 10 min to 100 h). The XLS-R 2B model gave the best performance when fine-tuned on 50 h of Maltese speech, showing 8.53% WER and 1.93% CER on the CommonVoice test set, and 24.98% WER and 8.37% CER on the MASRI test set. Comparable results were achieved with fine-tuning Wav2Vec 2.0 models on at least 10 h of speech. Whisper models showed worse results, especially Whisper-tiny model, which resulted in the value of WER equal to 100% for all experiments.

Another research on comparison of Wav2Vec 2.0 and Whisper was presented in [14], where authors describe fine-tuning procedures for Wav2Vec 2.0 Base and XLSR-53, as well as for Whisper Small and Large models in the context of the Kazakh speech

recognition. According to this paper, Wav2Vec 2.0 outperformed Whisper-based models, with the values of WER equal to 9.8% and CER equal to 2.7%. Additionally, Wav2Vec 2.0 Base model was contrast to Multilingual Whisper Large model, which gave 19.8% WER and 4.1% CER.

However, some techniques of fine-tuning and optimization of Whisper, such as distillation [15], may prove successful in competitive results [16] on low-resource ASR tasks. For example, the best results for the North Sámi language, a low-resource language which is relative to Karelian, were achieved using Whisper large [17] and amounted to 24.91% WER on 34 h of speech.

Following this trend, it is clear that training an ASR system for a low-resource language such as Livvi-Karelian should leverage these larger models. Thus, one of the main aims of the present research is to explore fine-tuning Wav2Vec XLS-R variants (300 m, MMS, w2v-BERT) and Whisper variants (small, medium, distill-large v2) for the purposes of development of a Livvi-Karelian ASR and to compare these models with a hybrid Kaldi-based model. Besides, this article also discusses such procedures as adding a N-gram language model and expanding the dataset by speech augmentation through pitch and tempo change.

2 Data Collection and Annotation

Radio broadcasts in Karelian were the only source for AnKaS, featuring interviews with Livvi-Karelian native speakers. A set of 13 broadcasts was selected, each following an interview format involving at least two speakers (an interviewer and an interviewee). Some broadcasts presented more than two speakers, and some of the interviewers took part in several broadcasts. Therefore, the speech corpus comprised 17 speakers: 7 men and 10 women. The audio recordings were transcribed by specialists in Livvi-Karelian, and further the annotation of the corpus was carried out, and the audio data were segmented into phrases (Fig. 1).

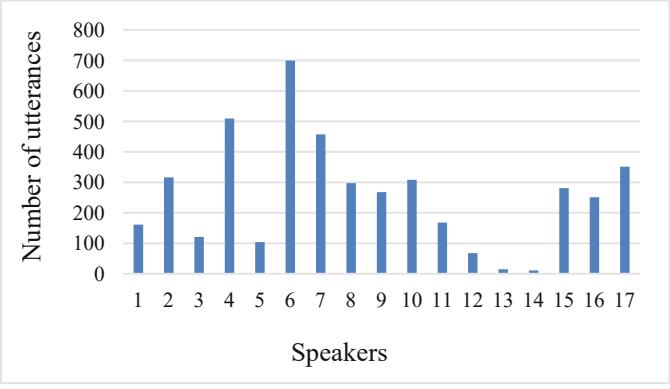


Fig. 1. Schematic diagram of speech production.

Any speech segments unsuitable for further analysis were deleted from the dataset (see below). The final corpus comprised 4.5 h of speech data, encompassing a total of 4385 recorded utterances, as illustrated in Table 1 below.

Table 1. Corpus metadata.

Corpus features	Value
Speakers	17 (7 male, 10 female)
Duration	4.5 h
Utterances	4385
Word occurences	32,037
Unique words	9117

The authors of this paper would like to emphasize the fact that the database presented in this work is not an audio corpus, but rather annotated transcripts (including segmentation into utterances, time- and code-switching tags). The audio data themselves are the property of The Russian Television and Radio Broadcasting Company (RTR). Therefore, the presented data do not in any way violate the rights of the copyright owner. However, external links to audio files are included in the annotation for the user's convenience.

The database is represented in JSON format. A separate.json file was created for each speaker. The following keys were used:

- "phrase_id" is the phrase number for this speaker
- "link" is a link to the audio recording
- "time_start" is start time of the phrase
- "time_end" is end time of the phrase
- "sentence" is textual transcription
- "sentence_rus" is textual transcription with code-switching, indicated with brackets and the tag "rus"

An example of annotation is given in Fig. 2.


```
{
    "phrase_id": "0002",
    "link": "https://tv-karelia.ru/wp-content/uploads/2022/01/Kirvesmies-Aleksandr-Ivanov.mp3",
    "time_start": "114.7394720",
    "time_end": "120.5044720",
    "sentence": "vot d'ad'a miša jakovlev täs susiedu d'ad'a pet'a jakovlev täs test'u opasti",
    "sentence_rus": "<vot d'ad'a miša jakovlev>rus täs susiedu <d'ad'a pet'a jakovlev>rus täs test'u opasti"
},
```


Fig. 2. An example of annotation entry.

The collected speech data revealed several issues that complicated the work and resulted in the removal of some entries from the final corpus. One of the primary issues encountered during the processing of recordings was simultaneous speech by several speakers, interrupting each other or speaking at the same time. Overlapping speech is difficult to process, and removing such segments is a non-trivial task, that is why utterances containing simultaneous speech were not included in the corpus.

Another factor that complicated the creation of the audio corpus was background noise. Despite using only studio-quality recordings, in some cases, background noise included music, the sound of page-turning, and (in one recording) traffic and street noises. Thus, all recordings containing background noise were also removed from the database.

Another issue was the so-called code-switching between Karelian and Russian. Code-switching ([18–21]) is a common phenomenon in the contemporary Karelian speech. The cases of code-switching encountered in the collected data are not numerous from the statistical point of view. For example, only 1.13% lexical units were marked with code-switching tags (366 word occurrences of total 32,037). However, given that the only source for the language data were radio broadcasts and not everyday communications, this number is still significant.

The statistics of the observed cases of code-switching is presented in Table 2.

Table 2. Code-switching statistics (lexemes grouped by lexical classes).

Lexical class	Lexemes
Nouns, Noun Phrases	42
Proper Names	66
Verbs	9
Interjections, Adverbs and Adverbials	38
Numerals	11

A linguistic commentary on the code-switching phenomena lies outside the scope of the present work, but it is important to recognize that the copied elements, especially those borrowed from the Russian language, frequently exhibit phonological characteristics that align with the Russian phonology. For instance, when incorporating Russian proper or common nouns, Karelian speakers may adapt the pronunciation to conform to Russian phonological norms. This adaptation may involve modifications in vowel quality, stress patterns, and consonant articulation to make the borrowed elements sound more congruent with the Russian phonology.

In order to perform evaluation of the collected corpus for speech recognition task, the data was divided into training, development, and testing parts, namely, 80% of utterances were used for acoustic model training, 10% were used for fine-tuning of the hyper-parameters of the model, and 10% were used for final testing. The size of the training part was enlarged by augmentation procedures made with the help of SoX toolkit. The modification of pitch on the number of semitones obtained randomly from

uniform distribution in range [-2, 2] was carried out. The speech rate perturbation was performed by coefficient randomly chosen from a uniform distribution in the range of [0.7, 1.3]. Additionally, simultaneous modification of both pitch and speech rate was applied. As a result, the size of the training set was enlarged up to 13.5 h.

3 Speech Recognition Experiments

3.1 Kaldi-Based Experimental Setup

At first, baseline experiments on Karelian speech recognition using the Kaldi toolkit [22] were conducted. Kaldi s5c recipe for chain model was applied for the training of acoustic models. For acoustic modeling a model similar to presented in paper [23] was used. Hybrid DNN/HMMs, based on factorized time-delay neural networks (TDNN-F) [24] was applied with Mel-frequency cepstral coefficients (MFCCs), complemented by an additional 100-dimensional i-Vector [25] being used as input features. DNN consisted of 16 layers. The first layer was TDNN layer, followed by three TDNN-F layers with time context of {-1, 0, 1}. The next layer was TDNN-F with no splicing. Then ten TDNN-F layers with time context of {-3, 0, 3} followed. Each TDNN-F layer had a dimension of 1024, with a bottleneck of 128.

The system's vocabulary with phonemic transcriptions were made automatically. For the Karelian language, the process of automatic transcriptions development was relatively straightforward, since the Karelian language features a fixed stress, consistently falling on the first syllable, and its vowels are less prone to reduction. As a result, the automatic transcription process primarily involves locating stress, identifying doubled graphemes as representations of long phonemes, and determining palatalized consonants (preceding front-row vowels). The process of the annotation development is described in detail in [26]. Speech decoding was carried out using a trigram language model trained with the help of SRI Language Modeling Toolkit (SRILM) [27]. Due to code-switching issues, the transcription dictionary includes Russian words as well. More details of Russian phonemic transcription issues were discussed in the previous works [28]. The N-best list rescoring was performed using neural network (NN) based language model trained with the use of TheanoLM toolkit [29]. The NN-based LM consisted of the projection layer with the size of 500 and two LSTM layers with the size of 512. Optimization criteria was Nesterov Momentum. Batch size was equal to 16. The LSTM-based model was linearly interpolated with the trigram model. For training both language models textual data was used, acquired from an open corpus of Veps and Karelian languages "VepKar", and from publications and journals in Livvi-Karelian, along with transcripts of the training part of the corpus. The details of textual data processing and language model are described in [23]. The text corpus was used for forming the system's vocabulary, which included all the words from the transcriptions of the training part of speech corpus and words from other text material that appeared in it at least twice. The size of ASR system's vocabulary was 143,907 words.

3.2 Wav2Vec-Based Experimental Setup

During Wav2Vec-based experiments, the training was conducted with the use of the Transformers framework [30]. The following pre-trained wav2vec models, briefly

described in the following paragraphs, were fine-tuned: Wav2Vec2.0-large-uralic-voxpopuli-v2, mms-1b-all, w2v2-bert.

Wav2Vec2.0-large-uralic-voxpopuli-v2 is a 300 m parameters variant of XLSR [5] acoustic model developed by Facebook AI Research, pre-trained on 42.5 h of unlabeled speech data of the Uralic languages from the VoxPopuli corpus [31]. Basically, it is an extension of the cross-lingual language model XLM-R, designed to handle multilingual and cross-lingual Natural Language Processing (NLP) tasks. Although XLSR-53 is built on the Wav2Vec 2.0 model, it can learn latent quantization that is spread across language. XLSR-53 uses product quantization to select quantized representations from codebooks, which are further selected using the Gumbel-Softmax method in a completely distinguishable manner. XLSR's architecture is similar to that of BERT [9]. XLSR's ability to understand multiple languages makes it particularly useful for cross-lingual transfer learning, where a model trained on one language can be adapted to another language with minimal additional training. This model was fine-tuned on the train dataset for 10 k steps, with batch size equals to 8 and 4 gradient accumulation steps.

mms-1b-all [7] is a model based on the Wav2Vec2 XLS-R [6] architecture. That model makes use of adapter models to transcribe 1000 + languages. This variant consists of 1 billion parameters and has been fine-tuned from facebook/mms-1b on 1162 languages, one of which was Karelian. Only an adapter model for Karelian was fine-tuned on the dataset in this case. That lowered the amount of trained weights, thus boosting the training process, and allowed not to lose trained information concerning the other languages. This model was fine-tuned on the train dataset for 10 k steps, with batch size equals to 8 and 4 gradient accumulation steps.

w2v2-bert is a method that combines the core methodologies from wav2vec 2.0 and BERT. The idea of w2v-BERT is to use the contrastive task defined in wav2vec 2.0 in order to obtain an inventory of a finite set of discriminative, discretized speech units. Further, these units are used as target in a masked prediction task in a way that is similar to masked language modeling proposed in BERT for contextualized speech representations learning. In the present work, v2 w2v-BERT 2.0 [11] was used, which comprises 24 Conformer layers [32] with approximately 600 M parameters and the same pre-training hyperparameters as v1. This model was pre-trained on 4.5 M hours of unlabeled data by the authors. Then, this model was fine-tuned on the train dataset for 10 k steps, with batch size equals to 2 and 16 gradient accumulation steps.

Furthermore, the impact of augmentation of the Wav2Vec 2.0 and MMS models with a 3-g language model trained on text data of the corpus was investigated. Previously it was shown that even a small language model can drastically improve the results of ASR for low-resource languages with code switching [33].

Unlike in the Kaldi-based approach for tokenization in Wav2Vec-based approach, a shortened vocabulary was used, which does not consider the length of vowels and consonants of Livvi-Karelian. The usage of all sounds was studied, however, it had no considerable impact on the training results. Furthermore, previous research of ASR development using end-to-end models for Finnish was not considering length of vowels [34], even though length of vowels and consonants is important for Finnish as well as for Livvi-Karelian. Also, the punctuation mark «‘» left in vocabulary separately of consonants after which it is usually placed, thus shortening the vocabulary.

3.3 Whisper-Based Experimental Setup

Overall, 3 versions of Whisper [12]: small, medium, and distilled large v2 [15] were explored. Whisper small has 244 M trainable parameters, Whisper medium has 769 M parameters, and distilled large has 756 M parameters. The usage of original large v2 and v3 versions was considered, however, due to high resource requirements they were not studied within this work. Instead, a distilled version of large model was used, which are said to be faster in exploitation and training, yet to lose only 1% on WER.

While working on tokenization and fine-tuning of Whisper-based models for Livvi-Karelian, which is out-of-boundaries of languages presented in the currently supported list, it was decided to fine-tune the model marking the task as ASR for Finnish, which is listed in the Whisper language list. This allowed reusing weights of the Finnish, which is closely related to Livvi-Karelian, wiping out Finnish in the process but providing ASR for Livvi-Karelian in the same way as it was done in [17] for North Sámi.

3.4 Speech Recognition Results

According to Kaldi receipt, triphone acoustic models were trained, namely models trained using Linear Discriminant Analysis (LDA) and Maximum Likelihood Linear Transform (MLLT), as well as a model trained using Speaker Adaptive Training (SAT) and feature space Maximum Likelihood Linear Regression (fMLLR). The latter was utilized for generation of force-alignment for NN-based model training. The results of speech recognition experiments with different types of acoustic models on both development (Dev) and testing (Test) parts of Karelian speech corpus in terms of WER are presented in Table 3.

Table 3. Experimental results on Karelian speech recognition.

Type of AM	WER (%)	
	Dev	Test
Triphone	41.04	44.72
LDA + MLLT	37.86	42.29
fMLLR	37.14	39.49
TDNN-F/HMM	27.13	28.77
TDNN-F/HMM + LSTM-based LM	25.44	27.20
Wav2Vec2.0-large-uralic-voxpopuli-v2	24.73	25.25
Wav2Vec2.0-large-uralic-voxpopuli-v2 + N-gram LM	19.69	19.83
mms-1b-all	31.56	32.06
mms-1b-all + N-gram LM	24.29	24.99
W2V2-bert	18.84	20.39

(*continued*)

Table 3. (*continued*)

Type of AM	WER (%)	
	Dev	Test
whisper-small	32.22	35.25
whisper-medium	25.54	28.54
distil-whisper-large-v2	28.38	30.75

The TDNN-F model allowed achievement of WER equal to 27.13% on the development set and 28.77% on the test set. The application of LSTM-based language model interpolated with trigram model with interpolation coefficient equal to 0.6 for N-best list rescoring gave the additional WER improvement. Thus, the best results with TDNN-F were the following: WER = 25.44% for the development set and WER = 27.20% for the test set. The bootstrapped confidence interval [35] computed with the help of Kaldi was [24.92, 29.47].

However, the usage of end-to-end Transformer-based, mostly Wav2Vec-based, models outperformed the results obtained with Kaldi. Fine-tune of Wav2Vec 2.0 has led to 25.25% WER on test set and 24.73% on development set. Augmentation with N-gram language model allowed achievement of even better results, thus showing 19.83% WER on test set and 19.69% on development set.

Fine-tune of the MMS model showed competitive, but not outstanding results, even though this model already had weights for Karelian ASR. Results were following: 32.06 WER on the test set and 31.56% on the development set. Addition of N-gram model considerably improved results, thus leading to achievement of 24.99 WER on test set and 24.29% on development set.

The best results without usage of the N-gram model, 20.39% WER on the test set and 18.84% on the development set, were obtained with the w2v2-bert 2.0 model.

The best results using the Whisper-based model were obtained with the Whisper-medium model, and they were 28.54% and 25.54% WER on test and development set correspondingly. It looks like the best results using Whisper-based approach might be achieved using Whisper large v3 model, however high requirements for tuning of this model did not allow testing this assumption.

Overall, the best results of 19.83% WER on the test set and 19.69% on the development set were obtained with the usage of Wav2Vec 2.0 large and additional augmentation with the N-gram model. The results obtained are at the level of world results for other low-resource languages.

4 Conclusions

In the current stage of the present research two tasks were addressed, specifically, preparation of speech data in the Karelian language (transcripts and annotations), and recognition experiments. All the tasks were successfully solved.

However, it is evident that further work concerning collecting and processing Karelian language data is needed. The current total speech data size is relatively small, and there is a certain imbalance in speakers: there are more female speakers in the corpus than males (10 vs 7), and the recordings primarily feature speakers of middle and older age. Additionally, more data collected in the field is necessary to accurately reflect linguistic nuances, particularly code-switching, and consider these issues when creating automatic recognition systems of spoken Livvi-Karelian.

The reported WERs on the development and test sets demonstrate the effectiveness of the proposed approach. Achieving a WER of 19.69% on the development set and 19.83% on the test set with the Wav2Vec 2.0 large model is a decent result, considering the challenges associated with low-resource languages. While the reported results are promising, further analysis, including error pattern identification, will provide valuable insights into the system's strengths and drawbacks.

The results the presented work can find their applications in various fields, including speech-to-text transcription, language preservation, and human–computer interaction in minority language communities. Future research may involve exploring additional data augmentation techniques, refining and tuning language models, and extending the study to other low-resource languages in order to further validate the proposed approach.

Acknowledgements. This research was funded by the Russian Science Foundation, grant number 24-21-00276, https://rscf.ru/en/project/24-21-00276/

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Laakso, J., Sarhimaa, A., Åkermark, S.S., Toivanen, R.: Towards openly multilingual policies and practices: Assessing minority language maintenance in Europe. Multilingual Matters, Bristol (2016)
2. Joshi, P., Santy, S., Budhiraja, A., Bali, K., Choudhury, K.: The state and fate of linguistic diversity and inclusion in the NLP world. In: Proceedings of the LREC 2012 – 58th Annual Meeting of the Association for Computational Linguistics, pp. 6282–6293. Virtual Conference, July (2020)
3. Bender, E.M.: On achieving and evaluating language-independence in NLP. Linguis. Issues Lang. Technol. **6**(3), 1–26 (2011)
4. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: a framework for self-supervised learning of speech representations. Adv. Neural Inf. Process. Syst. **33**, 12449–12460 (2020)
5. Conneau, A., Baevski, A., Collobert, R., Mohamed, A., Auli, M.: Unsupervised cross-lingual representation learning for speech recognition. In: Proceedings of the Interspeech 2021, pp. 2426–2430 (2021)
6. Babu, A., et al.: XLS-R: self-supervised cross-lingual speech representation learning at scale. In: Proceedings of the Interspeech 2022, pp. 2278–2282 (2022)
7. Pratap, V., et al.: Scaling speech technology to 1,000+ languages. J. Mach. Learn. Res. **25**(97), 1–52 (2024)

8. Poth, C., et al.: Adapters: a unified library for parameter-efficient and modular transfer learning. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 149–160 (2023)
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
10. Chung, Y.A., et al.: W2v-BERT: combining contrastive learning and masked language modeling for self-supervised speech pre-training. In: 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 244–250 (2021)
11. Barrault, L., et al.: Seamless: Multilingual Expressive and Streaming Speech Translation. arXiv preprint arXiv:2312.05187 (2023)
12. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: Proceedings of the 40th International Conference on Machine Learning 2022 (ICML 2023), pp. 28492–28518 (2023)
13. Williams, A., Demarco, A., Borg, C.: The applicability of Wav2Vec 2.0 and Whisper for low-resource Maltese ASR. In: Proceedings of the 2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL 2023), pp. 39–43 (2023)
14. Kozhirbayev, Z.: Kazakh Speech Recognition: wav2vec2. 0 vs. Whisper. J. Adv. Inf. Technol. **14**(6), 1382–1389 (2023)
15. Gandhi, S., Platen, von, P., Rush, A.M.: Distil-Whisper: Robust knowledge distillation via large-scale pseudo labelling. arXiv preprint arXiv:2311.00430 (2023).
16. Liu, Y., Yang, X., Qu, D.: Exploration of Whisper fine-tuning strategies for low-resource ASR. EURASIP J. Audio, Speech Music Process. **2024**, 29 (2024). https://doi.org/10.1186/s13636-024-00349-3
17. Hiovain-Asikainen, K., De la Rosa, J.: Developing TTS and ASR for Lule and North Sámi languages. In: Proceedings of the 2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL 2023), pp. 48–52 (2023) https://doi.org/10.21437/SIGUL.2023-11
18. Heller, M.: Codeswitching: Anthropological and Sociolinguistic Perspectives. De Gruyter Mouton, Berlin – New York (1988)
19. Gardner-Chloros, P., Edwards, M.: Assumptions behind grammatical approaches to code-switching: when the blueprint is a red herring. Trans. Philol. Soc. **102**(1), 103–129 (2004)
20. Myers-Scotton, C.: Duelling languages. Grammatical structure in code-switching. Clarendon Press, Oxford (1993)
21. Sarhimaa, A.: Syntactic transfer, contact-induced change, and the evolution of bilingual mixed codes: Focus on Karelian-Russian language alternation. Finnish Literature Society, Helsinki (1999)
22. Povey, D., et al.: The Kaldi speech recognition toolkit. In: Proceedings of the ASRU 2011 – 2011 IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 1–4. Waikoloa, HI, USA (2011)
23. Kipyatkova I., Kagirov I.: Deep models for low-resourced speech recognition: Livvi-Karelian case. Mathematics **11**(18), ID 3814 (2023) https://doi.org/10.3390/math11183814
24. Povey, D., et al.: Semi-orthogonal low-rank matrix factorization for deep neural networks. In: Proceedings of the INTERSPEECH 2018 – 19th Annual Conference of the International Speech Communication Association, pp. 3743–3747. Hyderabad, India (2018)
25. Saon, G., Soltau, H., Nahamoo, D., Picheny, M.: Speaker adaptation of neural network acoustic models using i-Vectors. In: Proceedings of the ASRU 2013 – 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 55–59. Olomouc, Czech Republic, (2013)
26. Kipyatkova, I., Kagirov, I.: Phone durations modeling for Livvi-Karelian ASR. In: Proceedings of the 25th International Conference SPECOM 2023, Springer, Lecture Notes in Computer Science, vol. 14339, pp. 87–99. Dharwad, India (2023)

27. Stolcke, A., Zheng, J., Wang, W., Abrash, V.: SRILM at sixteen: update and outlook. In: Proceedings of ASRU 2011 – 2011 IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 5–9. Waikoloa, HI, USA (2011)
28. Karpov, A., Markov, K., Kipyatkova, I., Vazhenina, D., Ronzhin, A.: Large vocabulary Russian speech recognition using syntactico-statistical language modeling. Speech Commun. **56**, 213–228 (2014). https://doi.org/10.1016/j.specom.2013.07.004
29. Enarvi, S., Kurimo, M.: TheanoLM – an extensible toolkit for neural network language modeling. In: Proceedings of INTERSPEECH 2016 – 17th Annual Conference of the International Speech Communication Association, pp. 3052–3056. San Francisco, CA, USA (2016)
30. Wolf, T., et al.: Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771 (2019)
31. Wang, C., et al.: VoxPopuli: a large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Vol. 1: Long Papers), pp. 993–1003 (2021)
32. Gulati, A., et al.: Conformer: Convolution-augmented transformer for speech recognition. arXiv preprint arXiv:2005.08100 (2020)
33. Ogunremi, T., Manning, C.D., Jurafsky, D.: Multilingual self-supervised speech representations improve the speech recognition of low-resource African languages with codeswitching. arXiv preprint arXiv:2311.15077 (2023)
34. Grosz, T., Getman, Y., Al-Ghezi, R., Rouhe, A., Kurimo, M.: Investigating wav2vec2 context representations and the effects of fine-tuning, a case-study of a Finnish model. In: Proceedings of INTERSPEECH 2023, pp. 196–200 (2023) https://doi.org/10.21437/Interspeech.2023-837
35. Bisani M., Ney, H.: Bootstrap estimates for confidence intervals in ASR performance evaluation. In: Proceedings of ICASSP 2004 – 2044 IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. I–409. Montreal, Canada (2004)

Advances in OpenASR21 Evaluation with Increased Temporal Resolution for Speech Self-supervised Learning Models

Vishwa Gupta(✉)

Centre de Recherche Informatique de Montréal (CRIM), Quebec, Canada
vishwa.gupta@crim.ca

Abstract. The OpenASR21 evaluation consisted of speech recognition for low resource languages in 3 evaluation conditions: constrained, contrained plus, and unconstrained. In this paper we investigate the constrained plus condition. In the constrained plus condition, we can use any self supervised learning (SSL) model to reduce the word error rate (WER). The idea was to get good speech recognition accuracy with only 10 h of acoustic training data for the 15 low resource languages in OpenASR21.

In this paper, we show that we reduce WER for all the 15 languages when we increase the temporal resolution of feature parameters computed from the speech SSL models from 20 ms to 10 ms. The temporal resolution of the SSL models is in general 20 ms. This increase in temporal resolution is done without retraining the SSL models. The resulting feature parameters with increased temporal resolution lead to 3.9% average absolute reduction in WER (from 1.2% for Javanese to 7.8% for Amharic) for the development set of the 15 languages in the OpenASR21 evaluation. We also compare WER for 5 different pre-trained SSL models in the low resource OpenASR21 languages scenario.

Keywords: OpenASR21 · Low-resource · Speech recognition · SSL models · Temporal resolution

1 Introduction

The OpenASR21 (Open Automatic Speech Recognition 2021) Challenge set out to assess the state of the art of ASR technologies under low-resource language constraints [17]. The task consisted of performing ASR on audio datasets in up to 15 different low-resource languages and 3 languages with case sensitive scoring, to produce the recognized text. Ten languages were carried over from the OpenASR20 challenge [16], and five new languages were added. A case sensitive scoring was also added for three of these languages: Kazakh, Swahili and Tagalog.

In the constrained condition, only a 10-hour audio Build dataset for that language can be used for training acoustic models. Additional text data, either

A. Karpov and V. Delić (Eds.): SPECOM 2024, LNAI 15299, pp. 69–81, 2025.
https://doi.org/10.1007/978-3-031-77961-9_5

from the Build dataset or publicly available resources, can be used for training the language model. No pre-trained large acoustic models were allowed.

In the constrained plus condition, we could also use large pre-trained models to extract features to reduce the word error rate (WER). In this paper we address the issue of using self supervised learning (SSL) models to reduce the WER in a low resource scenario (10 h of labeled training audio).

A good overview of OpenASR20 is given in [16]. In OpenASR20, two teams achieved very good results [1,27]. They used larger training text and lexicon from Linguistic Data Consortium (LDC) corpora for training language models (LM) and using a larger lexicon. These LMs and larger lexicon reduced the WER significantly for each language.

For OpenASR21 see [17] for a good overview. The team from USTC/iFlytek Research [29] achieved the lowest WER for all the 15 languages in the constrained condition. Their WER was significantly lower than any other participant for all the languages. For acoustic modeling, they used text-to-speech (TTS) to generate additional audio for training either from public text or the Babel training text. This gave them an additional 1.3% average WER reduction for the 15 languages. They also rescored the decoded lattices with bidirectional LSTMP (LSTM with a recurrent projection layer) [21] language model from public text. Note that, the leading teams in the OpenASR21 evaluation used hybrid Deep Neural Network–Hidden Markov Model (DNN-HMM) systems rather than end-to-end systems, since the end-to-end systems perform poorly with only 10 h of audio.

In [7], the authors Improve Language Modeling, Voice Activity Detection, and Lexicon to reduce WER for the development (dev) set for the 15 languages in the constrained condition. They get lower WER for the dev set for 3 OpenASR21 languages than in the USTC/iFlytek system [29].

The team from THUEE [28] achieved the lowest WER for the eval set for 11 out of 15 OpenASR21 languages in the constrained plus condition (see Fig. 1 in [17]). They fine-tuned the XLSR-53 SSL model [5] with the training data for each language. They used two different ways of fine-tuning this model with the 10 h of training data for each language. In the first method, they fine tune the pre-trained XLSR-53 model with the labeled 10 h data of the target language by adding a linear classifier on top of the SSL model to optimize the CTC loss. The parameters of the feature extractor layers in XLSR-53 are not updated, but the parameters of the encoder layers and the classifier are updated. The WER for the dev set for a single decode with a 4 g language model can be seen in column FT of Table 4 in [28]. In the second method, they continue the self-supervised training of the XLSR-53 model with the unsupervised speech of the target language from the Babel audio (which is about 4 times the OpenASR21 training audio) and then fine-tune the resulting SSL model with the labeled 10 h audio as in the first method. Their lowest WER on the dev set for a single decode with this second method is in Table 4 column FT2 in [28]. All their experiments are run at a temporal resolution of 20 ms. We show that with just the feature parameters from the XLS_R-2b model (without any tuning of this model with

the target language training audio), at the increased temporal resolution of 10 ms, we get significantly lower WER than in the FT2 column of Table 4 in [28].

The reason we are just mentioning the THUEE system [28] for the constrained plus results is because they got the lowest WER for the evaluation (eval) set for 11 of the 15 languages in the constrained plus condition (see Fig. 1 in [17]), and Table 4 in [28] shows single decoding WER on the **dev** set for all the languages. We can use this Table to compare with our WER on the dev set. We also compare our results with the WER for the dev set in [29] for the constrained condition as they achieved lower WER on the dev set for 4 languages than the THUEE team in [28] for the constrained plus condition.

Recently, there has been some attempt to train multi-resolution speech SSL models instead of the typical 20 ms resolution for the SSL speech models. For example, in [25], the authors propose utilizing HuBERT representations at multiple resolutions for downstream tasks. They explore two approaches, parallel and hierarchical, for integrating HuBERT features with multiple resolutions. In their parallel approach, they combine 400, 100 and 20 ms temporal resolutions, but they do not go lower than 20 ms. Also in Table 4 in [25] we see that the error rate goes down as the frame interval goes from 400 to 100 to 20 ms, suggesting that an even smaller frame interval like 10 ms might produce even lower error rates.

In [24], the authors introduce a SSL model that leverages a hierarchical Transformer architecture, complemented by HuBERT-style masked prediction objectives, to process speech at multiple resolutions. They show improved results on LibriSpeech [13] and on Multilingual SUPERB data [23].

We also increased the resolution of the features obtained from the pre-trained SSL speech models from 20 ms to 10 ms. However, we do not resort to re-training the full pre-trained SSL model, which is computationally intensive and not all labs are equipped to do this kind of training. Instead, we extract features from the original audio, and from this audio trimmed by 10 ms in the beginning. We then combine the two sets of features by concatenating frames alternately to give a new set of features with 10 ms frame interval.

In this paper, we focus on using features extracted from large pre-trained SSL models at higher resolution (10 ms instead of 20 ms) to reduce WER for the 15 OpenASR21 languages in a very low resource environment. The experiments we ran to illustrate this used just the raw features from the pre-trained SSL models, and trained TDNN-F (factored time delay neural network) acoustic models using lattice free maximum mutual information (LF-MMI) [19] with these features. No discriminative training was carried out. Also, we used a simple 4-gram language model (LM) for decoding. The idea here is to show that just increasing the feature resolution from 20 ms to 10 ms results in a significant reduction in WER. This is not only true about the features from SSL models, but also for MFCC (Mel-frequency cepstral coefficients) features [6].

2 Dataset and Preprocessing

In this paper, for acoustic model training, we only used the 10-h Build data set provided by NIST for the language being processed, with corresponding transcripts in UTF-8 encoding. Training and development lexicons were also provided by NIST.

For the 13 languages with LDC packs (all the languages except Farsi and Somali), we used the expanded lexicon and text provided in those packs. For example, the training text in the OpenASR21 Build dataset varies from 66k words for Kazakh to 126k words for Vietnamese, while the training text in the LDC packs varies from 270k words for Kazakh to 989k words for Vietnamese. Overall, the LDC training text is between 4 times and 8 times larger than the text in the OpenASR21 Build. The lexicon in the LDC packs is also much larger than the lexicon in the OpenASR21 Build. For example, the number of words for Vietnamese in the OpenASR21 lexicon is 3.2k, while in the LDC lexicon there are 6.4k words. For these reasons, training a language model from the training text and lexicon in the LDC packs reduces the word error rate significantly for all the 13 languages with the LDC packs. In [28], the authors used the expanded training audio in these LDC packs to tune the XLSR-53 model for each language with self supervised learning. However, in this paper, we have strictly adhered to using 10 h of transcribed audio per language provided for the OpenASR21 training.

3 ASR Approach for Different Temporal Resolutions

The idea here is to compare features extracted from the SSL speech models with temporal resolution of 20 ms versus 10 ms. The features from the SSL models were extracted using Transformers[1]. These features are then used to train hybrid HMM-DNN systems based on WFSTs (Weighted Finite-State Transducers) using the Kaldi toolkit [18] (downstream fine-tuning). Separate DNN-HMM systems are trained for 20 ms and 10 ms frame intervals. Sec. 3.2 shows how to obtain features with 10 ms frame interval from the SSL speech models trained with 20 ms frame interval. Features with 10 ms frame interval are extracted without retraining the SSL models.

3.1 Training the Hybrid DNN-HMM System

For the DNN system, we train a factored time delay neural network (TDNN-F) [19]. Training the TDNN-F system requires generating alignment and lattices for the acoustic data. The alignment and lattices are generated using a GMM/HMM system trained with the 10 h of training audio files using 13-dimensional perceptually weighted linear prediction (PLP) features [10] (except for Cantonese and Vietnamese where we add 3 more pitch features). The HMM/GMM are computed separately for PLP features with 20 ms and 10 ms temporal resolutions.

[1] https://huggingface.co/docs/transformers/en/index.

Note that MFCC and PLP features with 20 ms and 10 ms temporal resolutions are computed with different window sizes in order to ensure reasonable speech overlap between consecutive windows. Window size for 20ms temporal resolution is 40 ms, while the window size for 10 ms temporal resolution is 25 ms. (Note that the feature extractor in the SSL models we have used computes the features from raw speech waveforms [3], so the issue of window size does not arise). We tried different model architectures for the two sets of features to get the lowest possible WER.

All the initial optimization experiments were run on Amharic data from OpenASR21 evaluation. Since the 20 ms frame interval leads to half the frames than with 10 ms frame interval, the optimized architectures for the two models turn out to be quite different.

The model for the 10 ms frame interval is shown in Fig. 1. We tried feature vectors generated by 5 different SSL models (XLSR-53 with 1024-dimensional features per frame and pre-trained on 56k hours of audio, XLS_R-300m with 1024-dimensional features per frame, XLS_R-2b with 1920-dimensional features per frame [2] (XLS_R models are pre-trained on 436k hours of audio), MMS-1b model [20] with 1280-dimensional features per frame and pre-trained on 491K hours of audio, and w2v-bert-2.0 pre-trained on 4.5M hours of audio (1024-dimensional features per frame, and 600 million parameters (see Sec 3.1 and Fig. 2 in [4])). Both the features from the SSL models and the i-vectors are input to the acoustic model. Using HMM/DNN based acoustic models allow us to use i-vectors [22] [8] together with the features from the SSL models. In the best scenario, the features from the SSL models and the 100-dimensional i-vectors are both transformed to 256 dimensions using a linear layer (see Fig. 1). The transformed 256-dimensional features from SSL model (after specAugment [14]) are then combined with the transformed 256-dimensional i-vectors as a single feature map using "combine-feature-map-layer" with a height of 128 and then passed through a single stream of 5 CNN layers similar to those in [9,12]. As in [12], we use 3×3 kernels for the 2D CNN layers with a filter size of 256 except for the first layer with a filter size of 128. We apply frequency band sub-sampling with a rate of 2 to every other layer in the 2D-CNNs. The output from the last CNN layer is fed to two streams of 256 dimensional TDNN-F layers with 13 TDNN-F layers in each stream. One stream has a time-stride of 3, and the other stream has a time-stride of 6. This model has 8.3 million parameters. The validation set for training uses 300 held-out utterances from the training set. Training time for these models for 10 epochs is around 17 h with 4 GPU's, and decoding time for 10 h of development set is 80 mins on 20 CPU's.

The optimized architecture for the acoustic model with a frame interval of 20 ms has 2 streams but only 4 TDNN-F layers in each stream. More than 4 TDNN-F layers significantly increased the word error rate (WER). Also, the "frame_subsampling_factor" of 3 for training the TDNN-F models was changed from 3 to 1 (as the features have a temporal resolution of 20 ms instead of 10 ms), and the time-strides of 3 and 6 for the two streams were changed to 2 and

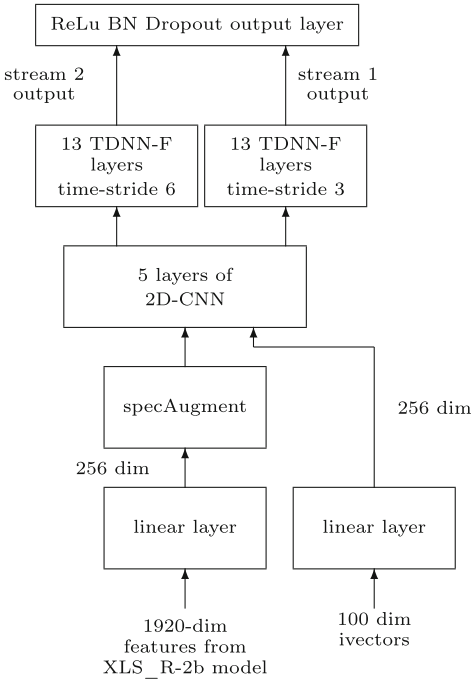


Fig. 1. Multi-stream TDNN-F acoustic model with 1920-dimensional features from 47th layer of XLS_R-2b model as input, together with 100 dimensional i-vectors.

4. This model has 7.5 million parameters. Training time for 10 epochs is 100 mins on 4 GPU's, and decoding 10 h of dev audio takes 30 mins on 20 CPU's.

3.2 Generating Features with 10 ms Frame Interval from SSL Models Trained with 20 ms Frame Interval

How can we generate features with 10 ms frame interval from any SSL model trained with 20 ms frame interval without having to re-train the SSL model with 10 ms frame interval? The answer is simple. We extract features from the SSL model with 20 ms frame interval using the original audio. We then trim the original audio by 10 ms in the beginning. The 20 ms frame-interval features from this trimmed audio are now shifted by 10 ms compared to features from the original audio. We then combine the two sets of features with 20-ms frame interval to generate features with 10-ms frame interval. The first feature frame comes from the first frame of the original audio. The second feature frame comes from the first frame of the trimmed audio. The 3rd frame comes from the 2nd frame of the original audio, and so on (see Fig. 2).

Note that, with a similar process, we can generate features with any given frame interval from an acoustic model trained with 20 ms frame interval. For example, to generate features with 5 ms frame interval, we trim the original

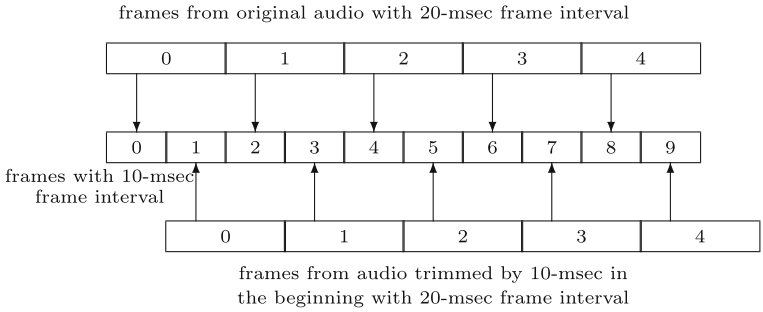


Fig. 2. How frames from original audio (top) and from audio trimmed by 10 ms in the beginning (bottom) are combined to give frames with 10-ms frame interval (middle).

audio by 5 ms, by 10 ms, and by 15 ms in the beginning and compute features for each of these trimmed audios. We then put together the frames of these four feature sets in proper order to generate features with 5 ms frame interval.

4 Initial Experiments with Different Temporal Resolutions

Initially, we ran all the experiments with different temporal resolutions on the Amharic language dataset from OpenASR21. For the pre-trained SSL model, we used XLS_R-300m model. We compared the word error rate (WER) on Amharic development (dev) set using TDNN-F acoustic model trained using the 1024-dimensional features from XLS_R-300m model (with frame interval of 20 ms and encoder layer 23) versus TDNN-F acoustic models trained with 40-dimensional MFCC features with a frame interval of 10 ms. We found that the WER with MFCC features (38.4%) was lower than that with features extracted from XLS_R-300m model (39.1%). We realized that some of the WER differences may be due to 10 ms frame interval for MFCC features versus 20 ms frame interval for features extracted from XLS_R-300m model. When we used MFCC features with frame interval of 20 ms (with window size of 40 ms), the lowest WER we could get was 51.3%. We quickly realized that we need to extract features with 10 ms frame interval from the existing SSL models trained with 20 ms frame interval. We simulated this with MFCC features first. We computed MFCC features at 20 ms frame interval with the original audio and then with each audio file trimmed by 10 ms. We then used the algorithm shown in Sect. 3.2 to compute MFCC features at 10 ms frame interval. With these simulated 10 ms MFCC features, we got 39.1% WER. This WER is a little bit higher than when we compute MFCCs directly at 10 ms because of the different window sizes (25 ms for 10 ms frame interval versus 40 ms for 20 ms frame interval). This experiment showed a large reduction in WER for MFCC features from 51.3% to 39.1% with simulated MFCC features with 10 ms frame interval. So we computed the feature parameters from XLS_R-300m at 10 ms frame interval

and we got a similar reduction in WER as shown in Table 1. Note that, we also computed WER with simulated 5 ms frame interval, but the WER was higher than with 10 ms frame interval for both MFCC features and features extracted from XLS_R-300m SSL models.

Table 1. WER for Amharic dev set with MFCC features and with features from 23rd encoder layer of XLS_R-300m model for frame intervals of 5 ms, 10 ms, and 20 ms.

Features	5 ms	10 ms	20 ms
40-dim MFCCs	39.9%	39.1%	51.3%
from XLS_R-300m model	33.4%	32.1%	39.1%

We compared five different SSL models: wav2vec2-large-xlsr-53, XLS_R-300m, XLS_R-2b, MMS-1b, and w2v-bert-2.0. All the five models gave similar results with the lowest WER for the w2v-bert-2.0 model (see Table 2). We then compared the WER for the dev sets for all the 15 languages in OpenASR21 evaluation for the best two models (w2v-bert-2.0 versus XLS_R-2b). It turns out that on average over 15 languages, the XLS_R-2b model gave 0.4% lower WER (42.3% versus 42.7%) than the w2v-bert-2.0 model (see Table 3). So the largest pre-trained SSL model XLS_R-2b gave the lowest WER.

Table 2. WER comparison for Amharic dev set with acoustic models trained with features from five different SSL models: wav2vec2-large-xlsr-53, XLS_R-300m, XLS_R-2b, MMS-1b, and w2v-bert-2.0.

Model	10 ms frame interval
wav2vec2-large-xlsr-53	34.9%
XLS_R-300m	32.1%
XLS_R-2b	31.1%
MMS-1b	31.6%
w2v-bert-2.0	**30.8%**

Since the XLS_R-2b model gave the lowest WER, we extracted feature parameters from 3 different encoder layers of this model to see if an intermediate layer will give lower WER. There was some issue as to which layer will give the best results [15]. We tried feature parameters extracted from layers 0, 35 and 47. The WER for the Amharic dev set with these feature parameters extracted at 10 ms frame interval is shown in Table 4.

As we can see from Table 4, the last encoder layer gives the lowest WER.

Table 3. Comparison of WER for dev sets from 15 OpenASR21 languages with features from XLS_R-2b model versus features from w2v-bert-2.0 model with frame interval of 10 ms.

Lang	XLS_R-2b	w2v-bert-2.0
Amharic	31.1	**30.8**
Cantonese	**41.8**	42.3
Farsi	**44.6**	46.9
Georgian	32.9	**32.3**
Guarani	**37.2**	38.1
Javanese	**45.7**	47.1
Kazakh	**38.6**	39.5
Kurmanji	**60.2**	61.3
Mongolian	**40.4**	40.6
Pashto	41.7	**41.4**
Somali	**53.9**	54.3
Swahili	30.4	**29.9**
Tagalog	**38.3**	39.4
Tamil	**57.1**	57.3
Vietnamese	40.0	40.0
Average	**42.3**	42.7

Table 4. WER for Amharic dev set with features from XLS_R-2b model from encoder layers 0, 35, and 47 at 10 ms frame interval.

Model	layer 0	layer 35	layer 47
XLS_R-2b	43.5%	31.7%	31.1%

5 Results and Comparison with Other Sites

Since the SSL model XLS_R-2b (layer 47) gave the lowest WER, we used the feature parameters from the 47th layer of this model to compare the word error rate (WER) for 10 ms frame interval vesus 20 ms frame interval. The idea was not to compare WER for different models, but to see how well features extracted from the SSL model with 10 ms frame interval compare against features with 20 ms frame interval and to compare our WER on the dev set for different languages with the published WER from other sites on the same task. We compare WER on the dev set because the NIST scoring server for the eval set is closed, and we do not have the transcripts for the eval set audio.

Table 5 compares our results with 10 ms frame interval (col 2), with 20 ms frame interval (col 3), with ref [29] (col 4), and with ref [28] (columns 5 and 6). Note that the results from [28,29] are not back-to-back comparable but are included for the reasons explained in the Introduction.

Table 5. Comparison of WER for dev set with features from XLS_R-2b model with frame interval of 10 ms (col 2), with frame interval of 20 ms (col 3), from ref [29] Table 1 col 2, from ref [28] Table 4 col FT (col 5) and col FT2 (col 6). Note columns 5 and 6 show **character** error rate for Cantonese (instead of WER).

Lang	10 ms	20 ms	from ref [29] table 1 column 2	from ref [28] table 4 column FT	from ref [28] table 4 column FT2
Amharic	**31.1**	38.9	35.0	44.0	38.6
Cantonese	41.8	43.8	42.3	37.0	36.6
Farsi	**44.6**	47.6	52.4	46.3	47.4
Georgian	**32.9**	34.5	37.5	44.7	47.3
Guarani	**37.2**	40.3	39.0	46.2	41.3
Javanese	**45.7**	46.9	51.9	53.6	49.8
Kazakh	**38.6**	43.3	46.1	46.5	42.1
Kurmanji	60.2	62.2	63.7	62.5	**59.5**
Mongolian	**40.4**	46.8	45.4	47.9	43.9
Pashto	41.7	45.9	45.4	45.2	**37.9**
Somali	**53.9**	58.7	55.9	**53.9**	55.8
Swahili	**30.4**	32.8	32.3	40.5	34.6
Tagalog	**38.3**	45.9	42.1	45.0	39.5
Tamil	**57.1**	63.5	61.0	64.3	60.0
Vietnamese	40.0	42.2	43.9	40.2	**36.6**
Average	**42.3**	46.2	46.2	47.9	44.7

As can be seen from Table 5, the WER with 20 ms frame interval is always higher than with 10 ms frame interval for every language. On an average, the WER with 10 ms frame interval is 3.9% absolute (or 8.4% relative) lower than with 20 ms frame interval. Also, note that for 11 of the 14 languages, the WER with 10 ms frame interval is lower than for any other column, and the average WER is lower by 2.4% absolute than the last column in Table 5. Both the comparisons are for a single decode of the dev set (without fusion from multiple decoders). These results show that we are getting state-of-the-art results on the OpenASR21 evaluation task.

Note that, the process of generating 10 ms frame interval features from 20 ms frame interval features can also be considered as a feature augmentation. In general, feature augmentation means perturbing the audio by some means. Some of these perturbations used frequently are speed perturbation [11], and adding MUSAN noise [26] randomly to audio samples. The generation of 10 ms features from 20 ms features can be considered as another kind of feature augmentation. We have shown in this paper that this kind of feature augmentation for features from SSL models results in significant reduction in WER for the low resource OpenASR21 languages.

6 Conclusion

In this paper, we show that extracting feature parameters at 10 ms frame interval from pre-trained SSL models in low resource scenario results in lower word error rate (WER) than with feature parameters extracted at the usual 20 ms frame interval. We show it in the low resource scenario of OpenASR21 evaluation where only 10 h of training audio is provided for 15 different low resource languages. For all the 15 languages, the WER with 10 ms frame interval is on an average 3.9% absolute lower (42.3% versus 46.2%) than with frame interval of 20 ms. Also, for every language, the WER with 10 ms frame interval is lower than with 20 ms frame interval.

Note that extracting feature parameters at 10 ms frame interval from pre-trained SSL models with 20 ms frame interval can also be considered as a new kind of feature augmentation.

We compare our results with the best teams in OpenASR21 evaluation for constrained and constrained plus conditions, and show that our results are state-of-the-art results on the 15 languages in OpenASR21 evaluation in the constrained plus condition.

We can compute feature parameters from the speech SSL models at any temporal resolution (or frame interval) without having to retrain the speech SSL model. So we can benefit from feature parameters at any temporal resolution from the SSL model depending on the application without any expensive retraining of the SSL models.

We also compare 5 different pre-trained SSL models and show that the XLS_R-2b model gives the lowest WER in the low resource OpenASR21 languages scenario.

Acknowledgments. The authors would like to thank Ministry of Economy and Innovation (MEI) of the Government of Quebec for the continued support.

References

1. Alumäe, T., Kong, J.: Combining hybrid and end-to-end approaches for the OpenASR20 challenge. In: Proceedings of the Interspeech, pp. 4349–4353 (2021)
2. Babu, A., et al.: XLS-R: self-supervised cross-lingual speech representation learning at scale (2021)
3. Baevski, A., Zhou, H., Mohamed, A., Auli, M.: wav2vec 2.0: a framework for self-supervised learning of speech representations (2020)
4. Communication, S., et al.: Seamless: multilingual expressive and streaming speech translation (2023)
5. Conneau, A., Baevski, A., Collobert, R., Mohamed, A., Auli, M.: Unsupervised cross-lingual representation learning for speech recognition. arXiv eprint arXiv:2006.13979 (2020)
6. Davis, S., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Process. **28**(4), 357–366 (1980). https://doi.org/10.1109/TASSP.1980.1163420

7. Gupta, V., Boulianne, G.: Improvements in language modeling, voice activity detection, and lexicon in OpenASR21 low resource languages. In: Karpov, A., Samudravijaya, K., Deepak, K.T., Hegde, R.M., Agrawal, S.S., Prasanna, S.R.M. (eds.) Speech and Computer, pp. 73–86. Springer Nature Switzerland, Cham (2023)
8. Gupta, V., Kenny, P., Ouellet, P., Stafylakis, T.: I-vector-based speaker adaptation of deep neural networks for French broadcast audio transcription. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6334–6338. IEEE (2014)
9. Han, K.J., Pan, J., Tadala, V.K.N., Ma, T., Povey, D.: Multistream CNN for robust acoustic modeling. In: Proceedings of the ICASSP, pp. 6873–6877 (2021)
10. Hermansky, H.: Perceptual linear predictive (PLP) analysis of speech. J. Acoust. Soc. Am. **87**(4), 1738–1752 (1990)
11. Ko, T., Peddinti, V., Povey, D., Khudanpur, S.: Audio augmentation for speech recognition. In: Proceedings of the Interspeech 2015, pp. 3586–3589 (2015). https://doi.org/10.21437/Interspeech.2015-711
12. Pan, J., Shapiro, J., Wohlwend, J., Han, K.J., Lei, T., Ma, T.: ASAPP-ASR: multistream CNN and self-attentive SRU for SOTA speech recognition. arXiv preprint arXiv:2005.10469 (2020)
13. Panayotov, V., Chen, G., Povey, D., Khudanpur, S.: LibriSpeech: an ASR corpus based on public domain audio books. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5206–5210 (2015). https://doi.org/10.1109/ICASSP.2015.7178964
14. Park, D.S., et al.: SpecAugment: a simple data augmentation method for automatic speech recognition. In: Proceedings of the Interspeech, pp. 2613–2617 (2019)
15. Pasad, A., Chou, J.C., Livescu, K.: Layer-wise analysis of a self-supervised speech representation model (2022)
16. Peterson, K., Tong, A., Yu, Y.: OpenASR20: an open challenge for automatic speech recognition of conversational telephone speech in low-resource languages. In: Proceedings of the Interspeech, pp. 4324–4328 (2021)
17. Peterson, K., Tong, A.N., Yu, J.: OpenASR21: the second open challenge for automatic speech recognition of low-resource languages. In: Proceedings of the Interspeech (2022)
18. Povey, D., et al.: The Kaldi speech recognition toolkit. In: Proceedings of the ASRU (2011)
19. Povey, D., et al.: Purely sequence-trained neural networks for ASR based on lattice-free MMI. In: Proceedings of the Interspeech, pp. 2751–2755 (2016)
20. Pratap, V., et al.: Scaling speech technology to 1,000+ languages (2023)
21. Sak, H., Senior, A.W., Beaufays, F.: Long short-term memory recurrent neural network architectures for large scale acoustic modeling. arXiv preprint arXiv:1402.1128 (2014)
22. Saon, G., Soltau, H., Nahamoo, D., Picheny, M.: Speaker adaptation of neural network acoustic models using i-vectors. In: 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 55–59 (2013). https://doi.org/10.1109/ASRU.2013.6707705
23. Shi, J., et al.: ML-SUPERB: multilingual speech universal performance benchmark (2023)
24. Shi, J., Inaguma, H., Ma, X., Kulikov, I., Sun, A.: Multi-resolution Hubert: multi-resolution speech self-supervised learning with masked unit prediction (2023)
25. Shi, J., Tang, Y., Inaguma, H., Gong, H., Pino, J., Watanabe, S.: Exploration on Hubert with multiple resolutions (2023)

26. Snyder, D., Chen, G., Povey, D.: MUSAN: a music, speech, and noise corpus (2015)
27. Zhao, J., et al.: The TNT team system descriptions of Cantonese and Mongolian for IARPA OpenASR20. In: Proceedings of the Interspeech, pp. 4344–4348 (2021)
28. Zhao, J., et al.: The THUEE system description for the IARPA OpenASR21 challenge. In: Proceedings of Interspeech, pp. 4855–4859 (2022). https://doi.org/10.21437/Interspeech.2022-649
29. Zhong, G., et al.: External text based data augmentation for low-resource speech recognition in the constrained condition of OpenASR21 challenge. In: Proceedings of the Interspeech, pp. 4860–4864 (2022). https://doi.org/10.21437/Interspeech.2022-649

Benchmarking Whisper Under Diverse Audio Transformations and Real-Time Constraints

Sergei Katkov(✉), Antonio Liotta, and Alessandro Vietti

Free University of Bozen-Bolzano, 39100 Bolzano, Italy
sergei.katkov@student.unibz.it

Abstract. The automatic speech recognition (ASR) domain has advanced considerably with the emergence of large transformer-based models, such as OpenAI's Whisper. This paper presents an experimental-based evaluation of the Whisper models, focusing on its performance under various acoustic conditions and input configurations. We specifically examine the effects of audio transformations such as white and Gaussian noise, reverberation, time stretch, and pitch shift, as well as the impact of varying chunk lengths. The findings suggest that while Whisper models are capable of dealing with minimal background noise and demonstrate commendable performance in clean audio conditions, their performance degrades rapidly when subjected to more severe audio transformations and noise, particularly when using shorter chunk lengths. This study contributes valuable insights into the Whisper model's capabilities and limitations, particularly when it comes to real-time speech recognition, offering guidance for future improvements in ASR technology.

Keyword: Automatic speech recognition.

1 Introduction

The accuracy of automatic speech recognition (ASR) models is crucial, as they frequently serve as the foundational layer for more complex systems. Recognizing the importance of this, the studies in [4,6] address the issue by adapting natural language processing (NLP) models to correct errors originating from ASR. Another method focuses on evaluating the robustness of ASR models [8], particularly when subjected to high levels of noise.

Recent years have seen significant progress in the field of ASR, particularly with the advent of large transformer-based models such as Whisper [15], wav2vec 2.0 [3], and Conformer [7] (to mention but a few).

OpenAI's Whisper models [15] are considered to be the state-of-the-art for speech-to-text conversion. This paper aims to present an in-depth evaluation of Whisper models, particularly focusing on their performance across various acoustic conditions and input configurations, which are crucial aspects for real-world applications. Additionally, this evaluation can also provide insights into

A. Karpov and V. Delić (Eds.): SPECOM 2024, LNAI 15299, pp. 82–91, 2025.
https://doi.org/10.1007/978-3-031-77961-9_6

pathological speech, where deviations from normal speech often resemble the transformations tested.

Our research primarily investigates the impact of different audio transformations—namely, white and Gaussian noise, reverberation, time stretch, and pitch shift—on the Whisper models' performance. These transformations are selected to mimic common auditory challenges appearing in real life and during online communication. We aim to assess the resilience and adaptability of Whisper models to such distortions. Additionally, this study examines the influence of the input sample chunk length on the model's accuracy, which is a critical parameter in real-time operations.

Furthermore, the Whisper models' ability to process multiple languages with high accuracy makes them an ideal candidate for our multilingual evaluation, encompassing English, Italian, and German. This approach not only tests the linguistic versatility of the model but also enhances the generalizability of our findings across different linguistic domains.

In conclusion, this paper seeks to contribute to the ASR field by systematically investigating Whisper models under varied acoustic and input conditions. The findings are intended to inform future developments in speech recognition technology, optimizing its application in diverse real-world scenarios and expanding its utility across multiple languages.

2 Related Work

The Whisper model, a significant development in ASR, was introduced in [15]. In this foundational paper, the authors extensively used the audio degradation toolbox [11] to evaluate the model's performance across various noise conditions. This approach was critical in establishing the robustness of the Whisper model in dealing with different types of noise, which is an essential factor in ASR technology.

However, the exploration of the Whisper model's performance under additional noise types and with limited input chunk lengths remains a less explored area in subsequent research. While the initial assessments in [15] set a benchmark for noise handling in ASR models, the specific challenges posed by a broader range of noises and shorter audio chunks have not been as thoroughly investigated in later studies.

It has been observed that Whisper-large models perform among the best across various languages, even under challenging noise and transformation conditions. In [9], it is shown that Whisper models maintain strong performance under various noise and transformation conditions across multiple languages.

The broader field of ASR research, has focused on creating noisy datasets [5] and developing noise augmentation techniques [1], contributing significantly to understanding and improving noise resilience in ASR models. However, these studies often do not directly address the Whisper model.

Additionally, it is important to clarify that the Whisper model is not inherently designed for offline applications, despite its heavyweight architecture and

preference for 30-s input chunks. This design choice, while making it more suitable for offline use, does not limit its potential for real-time applications. Research efforts to reduce the recognition latency of Whisper models, as mentioned in [10], are crucial in expanding its use cases, particularly for scenarios requiring immediate responses or involving other types of real-time constraints.

In summary, the Whisper model, introduced in [15] and tested against various noise conditions using the audio degradation toolbox, represents a major step forward in ASR technology. However, the model's performance in even more diverse noise environments and with shorter input lengths is an area ripe for further research, offering opportunities to enhance its capabilities and broaden its applications. This is, in fact, the main focus of our paper.

3 Methodology

The methodology of this study is anchored in simulating real-world audio conditions to evaluate the robustness and versatility of the Whisper models developed by OpenAI. In real-world scenarios, ASR systems are frequently challenged by a variety of audio disturbances and irregularities. Therefore, it is imperative to test these models under conditions that closely mimic such disturbances to assess their practical effectiveness.

To this end, our approach involves subjecting the Whisper models to a series of controlled audio transformations. These transformations are designed to replicate common acoustic challenges encountered in everyday environments. Such challenges include background noise in urban settings, distortions due to varying room acoustics, and fluctuations in speech quality due to different recording conditions. By systematically applying these transformations, we aim to closely observe how the Whisper models perform under stress and identify areas where their performance may be further improved for real-world application.

The choice of specific audio transformations is guided by both contemporary research in the field and the operational realities of ASR systems. The goal is to create a testbed that not only challenges the models but also remains grounded in practicality. This involves selecting transformations that are representative of typical scenarios in which ASR systems are deployed, ranging from noisy streets to indoor spaces with varying acoustic properties.

3.1 Audio Transformations

In assessing the performance of ASR models like Whisper under varying acoustic conditions, it is crucial to apply a range of audio transformations that can effectively simulate real-world auditory challenges. Each transformation selected for this study serves a specific purpose, replicating different types of disturbances that are commonly encountered in everyday audio environments. By introducing these elements into our test data, we can rigorously evaluate the resilience and adaptability of the Whisper models. This not only tests the models' ability to

maintain accuracy under stress but also provides insights into potential areas for improvement in handling diverse and challenging audio scenarios.

The audio transformations applied in this study include:

- **White Noise:** A signal characterized by equal power distribution across all frequencies, expressed as

$$n(t) = \alpha \cdot \text{rand}(t), \tag{1}$$

  where α denotes the amplitude, t represents time, and $\text{rand}(t)$ generates uniformly distributed random values.
- **Gaussian Noise:** Statistical noise following a normal distribution, given by

$$n(t) = \mathcal{N}(\mu, \sigma^2, t), \tag{2}$$

  where $\mathcal{N}(\mu, \sigma^2)$ represents a Gaussian distribution with mean μ and variance σ^2, and t is time.
- **Time Stretch:** Modifies the length of an audio signal while maintaining its original pitch, given by

$$y(t) = x(a \cdot t), \tag{3}$$

  where a is the stretch factor and t is the time variable.
- **Pitch Shift:** Changes the pitch of an audio signal using Fourier Transform techniques, expressed as

$$y(t) = F^{-1}\{F\{x(t)\} \cdot e^{j2\pi \Delta f t}\}, \tag{4}$$

  with Δf being the frequency shift and t representing time. In our experiments, we vary the parameter *n_steps*, which is directly proportional to Δf through the relation $\Delta f = n\_steps \times \frac{f_0}{12}$, where f_0 denotes the reference frequency prior to shifting. A shift of one *n_steps* results in a pitch change of one semitone.
- **Reverberation:** Imitates the effect of sound reflections, described by

$$y(t) = x(t) + \alpha \cdot x(t - \Delta t), \tag{5}$$

  where α denotes the decay factor, Δt is the delay time, and t stands for time.

Parameters for these transformations were chosen to ensure that sentence samples remain recognizable to human listeners, albeit with a noticeable level of noise or difficulty in recognition.

3.2 Chunk Length Variation

A critical aspect of our investigation involves analyzing the effect of varying chunk lengths on Whisper's performance. The original paper [15] states that audio files are split into 30-second chunks for further processing. We experimented with different chunk sizes to understand how they influence the accuracy and efficiency of speech recognition, particularly in scenarios with constrained computational resources and processing time. This exploration is particularly relevant for real-time applications.

3.3 Multilingual Analysis

To assess the versatility of Whisper models, we conducted our evaluation using the Common Voice dataset [2] in three languages: English, German, and Italian. We used multilingual models with passing language label. This multilingual approach allows us to explore the linguistic adaptability of the models under the applied audio transformations and chunk length variations.

4 Results

In this study, we assess speech recognition performance using the Word Error Rate (WER) metric. WER is calculated as:

$$\text{WER} = \frac{S + D + I}{N}, \tag{6}$$

where S, D, and I stand for the number of substitutions, deletions, and insertions needed to align the system's output with the reference transcription, and N is the total word count in the reference. Lower WER values signify better transcription accuracy.

Other than in the original paper [15], we do not perform advanced text normalisation prior to WER evaluation because the authors only provide their text normalisation approach for English; thus a comparison between different languages would not be correct. Instead, we only remove punctuation marks and other non alphanumeric symbols. All experiments are performed on the Common Voice 13.0 dataset on test-split for English, Italian and German languages, respectively.

A color map is utilized to visually represent the WER performance across various transformations and chunk lengths. The color scale transitions from green to orange and red, representing a progression from low to high WER values, respectively. Since the effects of different transformations vary in strength, the color map for each table is adjusted individually to better reflect the range of WER values observed in that specific context.

Table 1. Different Whisper models with various languages and chunk length, WER.

chunk_len	whisper-base			whisper-medium			whisper-large-v2		
	english	german	italian	english	german	italian	english	german	italian
2	0.91	1.47	1.87	0.58	0.88	0.78	0.57	0.85	0.74
4	0.49	0.62	0.87	0.24	0.26	0.29	0.23	0.21	0.25
8	0.30	0.33	0.42	0.14	0.11	0.12	0.13	0.09	0.10
16	0.26	0.30	0.37	0.13	0.09	0.11	0.11	0.07	0.08
30	0.26	0.30	0.37	0.13	0.09	0.11	0.11	0.07	0.08

In our research, the Whisper large v2 ASR model was subjected to a series of tests across various languages, with a particular focus on the influence of chunk length together with performing audio transformations.

The findings suggest that the German language model experiences a drastic decline in performance with variations in chunk length (Table 1), a phenomenon potentially related to the language structure and the average word length characteristic of German. Conversely, the English language model demonstrated superior resilience under analogous experimental conditions. This phenomenon persists on large and medium model sizes. The English language base model exhibits enhanced performance comparing to other base models.

We evaluated the whisper large v2 model for each language, with combinations of different grades of noise and chunk lengths. We chose noise parameters in such a way as to keep audio recognizable by a human: White Noise ($\alpha = 0.002$), Gaussian Noise (mean 0, std 0.002), Time Stretch ($\alpha = 0.8$), Pitch Shift (2 semitones), Reverberation (1.5 s). We provide the results in Table 2, 3, 4

Table 2. WER performance of whisper-large-v2 english model under various noise conditions.

chunk/noise	no noise	white	Gaussian	pitch-shift	time-stretch	reverb
2	0.57	0.64	0.63	1.01	1.14	0.99
4	0.23	0.27	0.27	0.53	0.65	0.67
8	0.13	0.16	0.16	0.31	0.35	0.44
16	0.11	0.14	0.14	0.29	0.27	0.39
30	0.11	0.14	0.14	0.29	0.26	0.39

Table 3. WER performance of whisper-large-v2 german model under various noise conditions.

chunk/noise	no noise	white	Gaussian	pitch-shift	time-stretch	reverb
2	0.85	1.02	1.04	1.66	2.07	1.24
4	0.21	0.30	0.30	0.66	0.83	0.66
8	0.09	0.14	0.13	0.31	0.33	0.35
16	0.07	0.11	0.11	0.28	0.25	0.31
30	0.07	0.11	0.11	0.28	0.25	0.31

Alterations in pitch, despite leaving the linguistic content of the audio unchanged, were observed to incur a significant degradation in the model's recognition capabilities. This outcome indicates that pitch variations present a substantial challenge to the ASR system, underscoring the sensitivity of the model to tonal changes in speech.

Table 4. WER performance of whisper-large-v2 italian models under various noise conditions.

chunk/noise	no noise	white	Gaussian	pitch-shift	time-stretch	reverb
2	0.74	0.82	0.81	1.36	1.52	1.10
4	0.25	0.26	0.27	0.66	0.84	0.61
8	0.10	0.11	0.10	0.32	0.38	0.31
16	0.08	0.09	0.09	0.29	0.26	0.27
30	0.08	0.09	0.09	0.29	0.26	0.27

Reverberation, introducing an echoic element to the audio, emerged as the most problematic auditory transformation for the ASR model. While human listeners readily adapt to and comprehend reverberated speech [13], the ASR model exhibited marked difficulties, suggesting an area for further development in echoic environment adaptation.

Interestingly, the incorporation of a modest degree of white noise into the audio environment yielded a negligible impact on the model's performance. This robustness against minimal background noise suggests potential applicability of the Whisper large v2 model in real-world scenarios where ambient noise is an unavoidable element.

In the following tables we test whisper-large-v2 model on the English language, varying noise/transformation parameters:

Table 5. Impact of white noise levels on WER across different chunk lengths.

Noise Level / Chunk Length	1	2	4	6	8	16	30
0.001	1.25	0.61	0.25	0.17	0.14	0.13	0.13
0.002	1.29	0.64	0.27	0.18	0.16	0.14	0.14
0.005	1.33	0.70	0.33	0.23	0.19	0.17	0.17
0.01	1.41	0.78	0.40	0.29	0.24	0.22	0.22
0.015	1.47	0.86	0.44	0.32	0.28	0.25	0.25
0.02	1.51	0.93	0.51	0.38	0.33	0.30	0.30
0.025	1.57	0.98	0.56	0.41	0.35	0.33	0.33
0.03	1.61	1.03	0.61	0.45	0.39	0.37	0.37

Incorporating white noise (Table 5) into the audio samples results in a progressive deterioration of the WER. At a noise level of 0.03, the transcription quality deteriorates significantly, sharply contrasting with the human listener's ability to effectively perceive and understand the audio under the same conditions [14].

The results concerning the impact of time stretching are presented in Table 6. It is observed that with lower stretch rates, the model's performance significantly

Table 6. Impact of time-stretch rate on WER for different chunk lengths.

Rate / Chunk Length	1	2	4	6	8	16	30
0.1x	9.43	4.77	2.88	2.73	2.98	3.29	2.22
0.3x	3.48	2.62	1.87	1.40	1.12	0.61	0.39
0.7x	1.86	1.26	0.72	0.49	0.37	0.26	0.26

deteriorates, particularly for shorter chunk lengths, which aligns with expectations. Interestingly, at a stretch rate of 0.7, where the audio remains fully comprehensible to human listeners [12], the model experiences a marked decline in recognition performance.

Table 7. Impact of reverberation time on WER across different chunk lengths.

Reverb Time / Chunk Length	1	2	4	6	8	16	30
1	1.72	1.06	0.69	0.50	0.41	0.38	0.38
1.5	1.82	0.99	0.67	0.51	0.44	0.39	0.39
2	1.77	0.97	0.62	0.45	0.39	0.36	0.36

The outcomes of introducing different reverberation times are detailed in Table 7. Given that reverberation introduces an echo effect with a specific delay, variations in the delay duration have a minimal impact on the WER. However, it is evident that the introduction of reverberation, regardless of the exact delay, substantially diminishes the recognition accuracy of the model.

Table 8. Impact of pitch shift on WER across different chunk lengths.

Steps/Chunk Length	1	2	4	6	8	16	30
1 step	1.53	0.97	0.53	0.37	0.31	0.28	0.28
2 steps	1.57	1.01	0.53	0.38	0.32	0.29	0.29
3 steps	1.58	1.00	0.58	0.38	0.32	0.29	0.29

The introduction of pitch shift to the audio samples (Table 8), as detailed in our findings, results in a notable degradation of the WER for speech recognition. However, the variance in WER across different pitch shift parameters is relatively insignificant.

The analysis of these results reveals that, although Whisper models demonstrate an ability to manage audio transformations and added noise, their performance is notably degraded if compared to a no-noise environment. The introduction of even light noise or subtle transformations results in a discernible

reduction in the models' effectiveness, diverging from their performance in clean audio conditions. This observed effect can be attributed to the model developers' approach of not employing any augmentations during the training process, which may have contributed to the model's reduced robustness to the specific transformations applied in our study.

5 Conclusion

The study conclusively demonstrates that the Whisper models, while robust in various aspects, exhibits certain limitations in handling audio transformations and noise. German language processing shows a comparably more marked decline in performance with changing chunk lengths in noise-free settings, likely due to the inherent structural complexity of the language. Pitch shifting and reverberation significantly impact the model's accuracy, underscoring its sensitivity to tonal and echoic changes. Overall, the performance of the models noticeably declines in comparison to their functioning in an environment without noise.

These results highlight areas for enhancement in the Whisper model, particularly in its adaptation to diverse acoustic conditions and its training process, which currently lacks augmentation. This research paves the way for further refinement of ASR systems, ensuring their applicability and efficiency across a wider range of real-world scenarios and languages.

Subsequent research should examine more sophisticated noise augmentation techniques and gain a deeper comprehension of the linguistic aspects that influence performance discrepancies. Furthermore, the acoustic transformations employed in this study may provide valuable insights for the enhancement of ASR systems for pathological speech, which often exhibits irregular pitch, breathiness, and other distortions. Addressing these aspects will result in ASR systems that are more inclusive and capable of serving users with diverse speech characteristics.

Additionally, it is crucial to investigate the discrepancies between human and machine intelligibility of distorted speech, with the aim of developing ASR systems that align more closely with human auditory perception. Such efforts will help develop ASR technologies that are robust, reliable, and effective in different linguistic contexts.

Acknowledgement. This work was carried out as part of the RATTLE (Voice Recogniser based on Artificial Intelligence) project, kindly funded by the Fondazione Pfizer.

References

1. Adolfi, F., Bowers, J.S., Poeppel, D.: Successes and critical failures of neural networks in capturing human-like speech recognition. Neural Netw. **162**(C), 199–211 (2023). https://doi.org/10.1016/j.neunet.2023.02.032
2. Ardila, R., et al.: Common voice: A massively-multilingual speech corpus. In: International Conference on Language Resources and Evaluation (2019). https://api.semanticscholar.org/CorpusID:209376338

3. Baevski, A., Zhou, H., Mohamed, A., Auli, M.: Wav2vec 2.0: a framework for self-supervised learning of speech representations. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS 2020, Curran Associates Inc., Red Hook, NY, USA (2020)
4. Cui, T., Xiao, J., Li, L., Jiang, X., Liu, Q.: An approach to improve robustness of NLP systems against ASR errors. ArXiv abs/2103.13610 (2021). https://api.semanticscholar.org/CorpusID:232352551
5. Duarte, J.C., Colcher, S.: Building a noisy audio dataset to evaluate machine learning approaches for automatic speech recognition systems. ArXiv abs/2110.01425 (2018). https://api.semanticscholar.org/CorpusID:238259030
6. Everson, K., et al.: Towards ASR robust spoken language understanding through in-context learning with word confusion networks. In: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 12856–12860 (2024). https://doi.org/10.1109/ICASSP48485.2024.10447938
7. Gulati, A., et al. (eds.): Conformer: Convolution-augmented Transformer for Speech Recognition (2020)
8. Higuchi, Y., Tawara, N., Ogawa, A., Iwata, T., Kobayashi, T., Ogawa, T.: Noise-robust attention learning for end-to-end speech recognition. In: 2020 28th European Signal Processing Conference (EUSIPCO), pp 311–315 (2021). https://doi.org/10.23919/Eusipco47968.2020.9287488
9. Katkov, S., Liotta, A., Vietti, A.: Robustness of ASR systems in multilingual and acoustically challenging environments. In: Proceedings of the 23rd International Conference of the Italian Association for Artificial Intelligence (AIxIA) (2024)
10. Macháček, D., Dabre, R., Bojar, O.: Turning whisper into real-time transcription system. In: Saha, S., Sujaini, H. (eds.) Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: System Demonstrations, 17–24. Association for Computational Linguistics, Bali, Indonesia (2023).https://aclanthology.org/2023.ijcnlp-demo.3
11. Mauch, M., Ewert, S.: The audio degradation toolbox and its application to robustness evaluation. In: International Society for Music Information Retrieval Conference (2013). https://api.semanticscholar.org/CorpusID:11675708
12. Müller, J.A., Wendt, D., Kollmeier, B., Debener, S., Brand, T.: Effect of speech rate on neural tracking of speech. Front. Psychol. **10** (2019). https://doi.org/10.3389/fpsyg.2019.00449, https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2019.00449
13. Nábělek, A.K., Robinson, P.K.: Monaural and binaural speech perception in reverberation for listeners of various ages. J. Acoust. Soc. Am. **71**(5), 1242–1248 (1982). https://doi.org/10.1121/1.387773
14. Payton, K.L., Uchanski, R.M., Braida, L.D.: Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing. J. Acoust. Soc. Am. **95**(3), 1581–1592 (1994). https://doi.org/10.1121/1.408545
15. Radford, A., Kim, J.W., Xu, T., Brockman, G., Mcleavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 202, pp. 28492–28518. PMLR (23–29 Jul 2023) (2023). https://proceedings.mlr.press/v202/radford23a.html

AutoMode-ASR: Learning to Select ASR Systems for Better Quality and Cost

Ahmet Gündüz(✉), Yunsu Kim, Kamer Ali Yuksel, Mohamed Al-Badrashiny, Thiago Castro Ferreira, and Hassan Sawaf

aiXplain Inc., San Jose, CA, USA
{ahmet,yunsu.kim,kamer,mohamed,thiago,hassan}@aixplain.com
https://aixplain.com/

Abstract. We present AutoMode-ASR, a novel framework that effectively integrates multiple ASR systems to enhance the overall transcription quality while optimizing cost. The idea is to train a decision model to select the optimal ASR system for each segment based solely on the audio input before running the systems. We achieve this by ensembling binary classifiers determining the preference between two systems. These classifiers are equipped with various features, such as audio embeddings, quality estimation, and signal properties. Additionally, we demonstrate how using a quality estimator can further improve performance with minimal cost increase. Experimental results show a relative reduction in WER of 16.2%, a cost saving of 65%, and a speed improvement of 75%, compared to using a single-best model for all segments. Our framework is compatible with commercial and open-source black-box ASR systems as it does not require changes in model codes.

Keywords: Automatic speech recognition · Quality estimation · Cost optimization

1 Introduction

Automatic speech recognition (ASR) has evolved remarkably due to advances in deep learning [5,13,14]. Consequently, numerous high-quality ASR models have been released [22,24,37], with some claiming to achieve human parity.

However, users frequently encounter challenges in selecting the most suitable ASR model for their speech data; the performance of various models can differ on the same segment, and their rankings may vary depending on the input conditions, such as accents, dialects, background noises, and speaking styles [4]. For instance, certain models are optimized for studio recordings, while others are more robust to non-speech noise. It is important to note that audio conditions can vary across different segments, even within the same corpus and application.

This variability poses a significant challenge in intelligently integrating multiple ASR systems for a specific purpose. Traditionally, this has been addressed through system combination [10,19], which constructs a confusion network from

A. Karpov and V. Delić (Eds.): SPECOM 2024, LNAI 15299, pp. 92–103, 2025.
https://doi.org/10.1007/978-3-031-77961-9_7

multiple hypotheses and finds the best path to derive the final transcription. Ensemble learning methods introduce diversity among the systems to expand the combination space [26,28]. Departing from the confusion network, [12] propose leveraging confidence scores from ASR systems to select the optimal hypothesis. While effective in reducing the Word Error Rate (WER), these approaches share a common limitation: they necessitate hypotheses from all candidate systems. Meeting this requirement is often impractical nowadays due to the large model size of high-performing modern ASR systems; commercial systems are expensive, and open-source models incur substantial costs.

This paper introduces AutoMode-ASR, a novel framework that predicts the most suitable ASR system for a given audio segment—defined as a contiguous chunk of speech, such as a sentence or phrase—without running the inference of candidate systems. The prediction is based on features extracted from the audio input, and its transcription is performed only with the predicted system afterward. The distinct separation between system selection and inference eliminates the requirement of modifying the decoding process, enabling a flexible combination of commercial and open-source models.

Our experimental results show that AutoMode-ASR improves transcription performance by up to -16.2% relative in WER. Notably, compared to other multi-system approaches, it does not increase operational costs; rather, it reduces costs by opting for a lighter system in cases of comparable performance, achieving a price reduction of 65%. Our contributions are:

- We present a new combination scheme for any ASR models at the segment level to optimize quality and cost.
- We analyze feature types to discern their relevance in accurately predicting the performance of ASR systems.
- We propose a robust classification module that facilitates the incremental integration of ASR systems.
- We demonstrate an effective approach to incorporate quality estimation for further optimizing performance.

2 Methodology

AutoMode-ASR aims to predict the optimal ASR system for a given audio input, a task framed as multi-class classification using system IDs as class labels. Training a classifier for this involves preparing data by conducting ASR inference for each candidate system on every audio segment, entailing significant costs. If thus only a limited amount of training data can be prepared, there is a high risk of insufficient cases for certain systems as top performers. This class imbalance adversely impacts classification accuracy [20,31], although it can be mitigated by data sampling [1,6] or boosting methods [29,34].

In this work, we approach the problem as learning to rank [18], leveraging the inherent rank information within our training data. Following the ASR inferences on a training segment, we acquire not only the ID of the top-performing

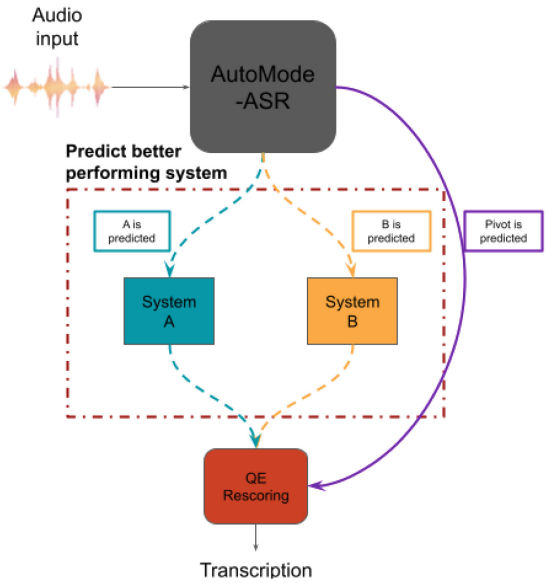


Fig. 1. The Diagram of the AutoMode-ASR Workflow.

system but also the rank across all candidates, sorted by WER. In contrast to classification, which is trained solely on the top system ID, ranking gives training instances for every system from each segment with relative performance. This maximizes the utilization of information within the training data and enhances stability.

2.1 Method Overview

Rather than directly predicting a ranking over all systems, we decompose the problem into multiple binary classification problems (Sect. 2.3), each comparing a specific pair of systems [15]. As mentioned, every training segment has ranking information between candidate systems. To train a binary classifier, we simply relabel each segment with the winning system between the two systems in question. Each binary classifier is then trained with all available segments, which is critical in our setup where data preparation is difficult.

Following the one-vs-one approach, learning to rank pairwisely ideally requires considering every pair of systems: $C(C-1)/2$ pairs with C as the number of systems. However, this proves impractical in numerous scenarios, leading to the development of various pair sampling methods [27,32]. In our method, we designate a *pivot* system from all candidates, playing the role of a comparative baseline against every other system. Each binary classifier compares the pivot against another candidate, resulting in $C-1$ pairs, referred to as *one-vs-pivot*. Among multiple systems, we strategically choose the cost-effective system as

the pivot. This decision introduces an implicit bias towards minimizing costs, aligning with our goal.

The decisions from individual binary classifiers are later merged into the final decision using a simple heuristic (Sect. 2.4). This two-pass strategy is advantageous when a new system is incrementally added to the comparison. In such cases, training another binary classifier between the pivot system and the new system suffices, which is significantly more efficient than retraining a multi-class classifier involving all systems. Figure 1 shows the entire workflow of the proposed method.

As our binary classifier, we employ the Gradient Boosting Machine (GBM) [11] due to its flexibility in feature integration and interpretability in analyzing feature importance. GBMs have demonstrated superiority over deep neural networks across various classification and ranking problems, excelling in terms of both accuracy and efficiency [9,25]. The GBM algorithm incrementally incorporates weak learners by fitting each new learner to the residuals of the preceding ensemble. Specifically, we utilize the eXtreme Gradient Boosting (XGBoost) package [7], where each weak learner is a regression tree.

2.2 Features

An essential research question in our work is identifying relevant aspects of audio input for predicting ASR performance. To study this question, we integrate diverse features in the binary classifiers designed to encapsulate audio files' acoustic, linguistic, and quality-related dimensions:

- **Self-supervised audio embedding**: Representations learned from speech audio via contrastive learning on the masked latent space prove beneficial for many downstream speech tasks [3]. We adopt the cross-lingual version of it, trained on 53 languages (Wav2Vec2-XLSR-53) [8]. We extract the last encoder states and average them over the time dimension to produce a consolidated vector of 1024 dimensions.
- **Input language**: AutoMode-ASR operates without assuming the audio input language; it can accommodate any supported languages across all systems. Recognizing that classifier decisions may differ based on language, we include the language of each speech segment as a categorical feature. This empowers the classifier to potentially select the model that performs better in that particular language.

In addition to extracting features directly from the audio file, we utilize a lightweight ASR model to capture valuable features from the inference process and its temporary transcription. Note that this ASR model differs from the systems compared within AutoMode-ASR. We opt for a compact and swift ASR model for feature extraction, thereby minimizing any significant increase in processing time. After the ASR inference, we obtain the following features:

- **ASR embedding**: We extract the output states of the encoder from the ASR model and compute their average over time. Our hypothesis posits that the

representations learned during transcription encapsulate precise information relevant to estimating transcription performance.
- **ASR confidence score**: ASR inference provides log probabilities for each output token, called confidence scores. We include the mean, standard deviation, and five-number summary of the probability values as features. These scores offer preliminary insights into the transcription's difficulty level; some systems may excel in deciphering ambiguous phonetics in challenging segments, while others may perform better in transcribing easier segments.
- **Quality estimation score**: Once we obtain a transcription from the feature-extracting ASR model, we can assess its quality even without a reference using quality estimation metrics; it serves as another indicator of speech recognizability. For this purpose, we utilize *NoRefER* [16,35,36], which computes a score for the transcription and exhibits a high correlation with WER. It is worth noting that the NoRefER score is calculated solely based on the transcription itself, without considering the audio, thus providing a distinct dimension of information compared to other features.
- **Quality estimation embedding**: NoRefER itself is a neural network, from which we can extract representations from its intermediate layers. Specifically, we extract the last hidden state before the final linear layer of the NoRefER network, resulting in an embedding of 384 dimensions.

2.3 Training

For training a binary classifier between the pivot and any ASR system, we first assign labels to each segment based on the system with the lower WER between the two. In cases where WER values are identical, we label them with the more cost-effective system. We observed numerous instances with identical WER values, particularly for short segments, resulting in substantial cost savings overall (Sect. 3.1).

To enhance training and increase the WER gain, we prioritize samples with carefully crafted sample weights, calculated as the product of the following factors:

- **Normalized WER difference**: Calculate the absolute WER difference between two systems and divide each value by the range of values across the entire training set. Segments with a larger difference in WER are given more weight in loss calculation as correctly classifying these segments is expected to yield greater gains.
- **Inverse label frequency**: To counteract bias toward a single system, we assigned higher weights to the minority label.

Classifier training minimizes the binary logistic loss along with its first and second-order gradients at each step of adding a weak regression tree [7]. The hyperparameters of the trees, such as the number of leaves, number of features, or learning rate, are selected using cost-frugal hyperparameter optimization [33], which samples a tree learner based on the estimated cost for improvement. Each hyperparameter setting is evaluated using cross-validation with WER reduction, i.e., WER decrease by AutoMode-ASR selections versus selecting a single system.

2.4 Multi-class Ensemble

For comparing all systems, we aggregate predictions from individual binary classifiers and choose the most confident decision. Every binary classifier provides a prediction and the probability between its two systems, which exceeds 0.5. We select the system predicted with the highest probability among multiple binary classifiers. Note that each binary classifier compares a system with the pivot system. If a system surpasses the pivot, it is selected; if none of the systems outperform the pivot, we default to the cost-effective pivot system.

For more elaborate decision-making and further cost savings, AutoMode-ASR offers an option to rescore comparisons when a more expensive system is chosen, i.e., the pivot is not selected. We obtain transcriptions from the systems and compute their quality estimation scores using NoRefER, ultimately selecting the system with the highest score. Since the comparison is based on system outputs and the WER-correlated NoRefER, we anticipate that this yields predictions more aligned with WER, while also providing another opportunity for the low-cost pivot to be selected. In contrast to the features involved in the initial decision (Sect. 2.2), this process requires running the ASR systems and comparing their transcriptions. The only extra cost is due to the NoRefER inference.

3 Experiments

For our experiments, we curated training and test data by selecting diverse audio samples from Common Voice [2] and LibriSpeech [21]. Combining these two sources, our dataset encompasses various speaking styles and recording acoustics. We included the English, French, Spanish, German, and Russian subsets from Common Voice to thoroughly evaluate AutoMode-ASR's adaptability and effectiveness across different languages. Each selected segment was then inputted into all systems under comparison to obtain the WER and the performance ranking. The statistics of the prepared data are in Table 1.

Table 1. The number of audio segments where each ASR system (System A, System B, System C, Whisper) achieved the best performance in terms of Word Error Rate (WER) in training, validation, and testing subset.

Top-Rank System	`train`	`valid`	`test`
System A	1,182	149	149
System B	1,322	189	174
System C	5,431	735	773
Whisper (pivot)	14,817	2,147	2,107
Total	22,752	3,220	3,203

We evaluated AutoMode-ASR using three commercial ASR systems (called System A/B/C[1]) alongside Whisper [24], an open-source model developed by OpenAI. These providers were selected due to their prominence in the industry and diverse speech recognition approaches, ensuring a comprehensive and practical evaluation. Whisper, specifically its "small" version, was chosen as the pivot among systems because of its low cost and latency. In feature extraction (Sect. 2.2), we employed the Whisper small model for ASR embedding and confidence score. While this choice coincides with the pivot model, it was made solely for our convenience and does not introduce any unintended bias toward the pivot in AutoMode-ASR.

To train the classifiers, we employed Microsoft's FLAML framework [30]. We conducted five-fold cross-validation with a time budget of 1,000 s for hyperparameter optimization (HPO). While we also involved LightGBM [17] and CatBoost [23] machines in the HPO process, XGBoost consistently outperformed the others; thus we only present the results obtained using XGBoost. The weighting scheme for training data sampling and the target metric for HPO are tuned according to the performance on a validation set.

System selection was assessed against the top-ranking systems using F1, weighted by inverse label frequency to address label imbalance (Table 1). Subsequently, ASR decoding was conducted with the selected system per segment, and WER was computed to evaluate AutoMode-ASR's actual improvement in transcription performance. These results were compared against selecting one system for all segments: the pivot system (pivot only), a non-pivot system (non-pivot only), or the system with the best overall performance when used for all segments (single-best). We removed punctuation marks and applied lowercasing before computing WER.

3.1 Main Results

Table 2 shows the reduction in WER and classification performance of each binary classifier. While the pivot system is competitive, it does not consistently outperform other commercial systems in all segments, particularly compared to System C. Even though Table 1 shows that the pivot outperforms System C in nearly three times as many segments, System C achieves a lower average WER than the pivot (Whisper), likely because System C excels in more challenging segments where WER reduction has a greater impact.

AutoMode-ASR's binary classifiers efficiently identify cases where alternative systems excel over the pivot, showcasing the framework's ability to optimize ASR system selection. Sample weighting in training (Sect. 2.3) consistently proves beneficial and QE rescoring provides an additional gain.

Table 3 displays the final results after ensembling all three binary classifiers. Compared to selecting a single system for all segments, AutoMode-ASR achieves significantly lower WER, decreasing from 13.4% to 11.6%, with QE rescoring

[1] The disclosure of the system providers is pending approval under legal review. Their names will be disclosed accordingly after the review.

further reducing it to 11.1%. Notably, this improvement does not increase operational costs or delays; it requires approximately 36% of the cost and 25% of the runtime of the single-best baseline. It is noteworthy that QE rescoring only introduces negligible extra cost and runtime. A small open-source model could nearly eliminate both the cost and runtime by selecting the pivot system exclusively. However, its performance is significantly inferior to that of AutoMode-ASR, as it does not benefit from strong commercial systems. Additionally, we provide the performance metrics when using actual top-performing systems ("Oracle"), indicating potential room for improvement in future work.

These figures not only highlight the system's processing efficiency but also its cost-effectiveness compared to the baseline'Single-best' system.

Table 2. Word Error Rate (WER) reduction and classification performance (F1-score) for each **binary classifier** comparing the pivot system (Whisper) against commercial systems (A, B, C). Results are shown for different system selection strategies, including AutoMode-ASR with and without sample weighting and quality estimation (QE) rescoring.

Pivot vs.	System A		System B		System C	
	WER [%]	F1 [%]	WER [%]	F1 [%]	WER [%]	F1 [%]
Non-pivot only	21.2	5.7	20.8	4.7	13.4	11.8
Pivot only	14.1	73.4	14.1	75.9	14.1	61.1
AutoMode-ASR	13.6	77.8	14.0	76.9	12.2	72.9
+ Sample weights	13.1	79.0	13.9	78.1	**11.3**	73.1
+ QE rescoring	**12.9**	**80.2**	**13.7**	**79.1**	11.4	**76.3**

Table 3. Multi-class ensemble results comparing Word Error Rate (WER), F1 score, cost, and runtime of different system selection strategies. The "Single-best" system represents the baseline. AutoMode-ASR and its variants with sample weighting and quality estimation (QE) rescoring achieve progressively lower WER at a reasonable decrease in cost and runtime. The "Oracle" represents the perfect prediction scenario.

System Selection	WER [%]	F1 [%]	Cost [%]	Runtime [%]
Single-best	13.4	9.4	100.0	100.0
Pivot only	14.1	52.2	**2.3**	**4.6**
AutoMode-ASR	12.3	62.5	18.6	19.3
+ Sample weights	11.6	63.4	36.2	24.9
+ QE rescoring	**11.1**	**65.5**	36.2	25.1
Oracle	6.5	100.0	41.0	37.4

3.2 Feature Ablation

Table 4 presents an ablation study about the impact of various features on predicting ASR performance, categorized into three groups: audio, ASR, and QE. All cases with QE features exhibit a clear improvement in WER compared to the case without. This underscores the value of QE scores and embeddings, which offer useful information distinct from audio or ASR features. Comparing the second and third rows reveals a similar effect between audio and ASR features.

Table 4. AutoMode-ASR results with different feature groups without QE rescoring, and when QE rescoring is applied to the best setting (all). "Audio" stands for self-supervised audio embeddings, "ASR" means ASR embedding and confidence scores, while "QE" includes QE score and its embedding. Input language feature is always used.

Feature Groups	WER [%]	F1 [%]
Audio + ASR	12.4	61.5
Audio + QE	11.8	61.5
ASR + QE	11.7	62.9
Audio + ASR + QE (all)	**11.6**	**63.4**
+ QE rescoring	11.1	65.5

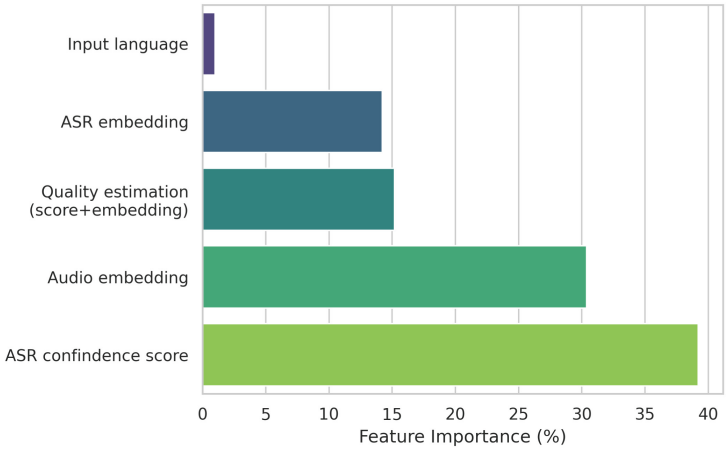


Fig. 2. Mean feature importance of binary classifiers.

The optimal configuration undoubtedly involves combining all features. Figure 2 illustrates the feature importance computed within the GBM. ASR confidence scores are deemed the most important, followed by embeddings from

self-supervised audio models, quality estimation models, and ASR models. This shows a considerable reliance on neural encoders for performance. Interestingly, language categorization appears to hold minimal importance, highlighting the AutoMode-ASR versatility across languages.

4 Conclusion

This work introduces AutoMode-ASR, a novel framework designed to dynamically select the most suitable ASR system for a given audio input; which harnesses the strengths of different ASR technologies to substantially improve transcription accuracy. It also considerably saves computational resources and operational costs by conducting binary system comparisons with a cost-effective system as the pivot. Through rigorous testing, AutoMode-ASR shows remarkable adaptability across audio environments and linguistic contexts, reducing WER from 13.4% to 11.1% with 65% lower cost and 75% faster speed. We verify that the multi-system ASR is a promising and practical way to optimize performance cost-effectively and time-efficiently through smart system selection.

References

1. Abdi, L., Hashemi, S.: To combat multi-class imbalanced problems by means of over-sampling techniques. IEEE Trans. Knowl. Data Eng. **28**(1) (2015)
2. Ardila, R., et al.: Common voice: a massively-multilingual speech corpus. In: Proceedings of the Twelfth Language Resources and Evaluation Conference, pp. 4218–4222 (2020)
3. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: a framework for self-supervised learning of speech representations. Adv. Neural Inform. Process. Syst. **33**, 12449–12460 (2020)
4. Benzeghiba, M., et al.: Automatic speech recognition and speech variability: a review. Speech Commun. **49**(10–11), 763–786 (2007)
5. Chan, W., Jaitly, N., Le, Q., Vinyals, O.: Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE (2016)
6. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. J. Artifi. Intell. Res. **16**, 321–357 (2002)
7. Chen, T., Guestrin, C.: Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016)
8. Conneau, A., Baevski, A., Collobert, R., Mohamed, A., Auli, M.: Unsupervised cross-lingual representation learning for speech recognition. arXiv preprint arXiv:2006.13979 (2020)
9. Ebrahimi, M., Mohammadi-Dehcheshmeh, M., Ebrahimie, E., Petrovski, K.R.: Comprehensive analysis of machine learning models for prediction of sub-clinical mastitis: deep learning and gradient-boosted trees outperform other models. Comput. Biol. Med. **114**, 103456 (2019)
10. Fiscus, J.G.: A post-processing system to yield reduced word error rates: recognizer output voting error reduction (rover). In: 1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings, pp. 347–354. IEEE (1997)

11. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Annals Stat. 1189–1232 (2001)
12. Gitman, I., Lavrukhin, V., Laptev, A., Ginsburg, B.: Confidence-based Ensembles of End-to-End Speech Recognition Models. In: Proc. INTERSPEECH 2023 (2023). https://doi.org/10.21437/Interspeech2023-1281
13. Graves, A., Mohamed, A.r., Hinton, G.: Speech recognition with deep recurrent neural networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE (2013)
14. Gulati, A., et al.: Conformer: convolution-augmented transformer for speech recognition. Interspeech (2020)
15. Hüllermeier, E., Fürnkranz, J., Cheng, W., Brinker, K.: Label ranking by learning pairwise preferences. Artif. Intell. **172**(16–17), 1897–1916 (2008)
16. Javadi, G., Yuksel, K.A., Kim, Y., Ferreira, T.C., Al-Badrashiny, M.: Word-level asr quality estimation for efficient corpus sampling and post-editing through analyzing attentions of a reference-free metric. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2024), Seoul, Korea, 14-19 April. IEEE (2024). https://doi.org/10.48550/arXiv.2401.11268
17. Ke, G., et al.: Lightgbm: a highly efficient gradient boosting decision tree. Adv. Neural Inform. Process. Syst. **30** (2017)
18. Liu, T.Y., et al.: Learning to rank for information retrieval. Foundat. Trends® Inform. Retrieval **3**(3) (2009)
19. Mangu, L., Brill, E., Stolcke, A.: Finding consensus in speech recognition: word error minimization and other applications of confusion networks. Comput. Speech Lang. **14**(4), 373–400 (2000)
20. Ou, G., Murphey, Y.L.: Multi-class pattern classification using neural networks. Pattern Recogn. **40**, 4–18 (2007)
21. Panayotov, V., Chen, G., Povey, D., Khudanpur, S.: Librispeech: an asr corpus based on public domain audio books. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5206–5210. IEEE (2015)
22. Pratap, V., et al.: Scaling speech technology to 1,000+ languages. arXiv preprint arXiv:2305.13516 (2023)
23. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A.: Catboost: unbiased boosting with categorical features. Adv. Neural Inform. Process. Syst. **31** (2018)
24. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: International Conference on Machine Learning, pp. 28492–28518. PMLR (2023)
25. Schmitt, M.: Deep learning vs. gradient boosting: Benchmarking state-of-the-art machine learning algorithms for credit scoring. arXiv preprint arXiv:2205.10535 (2022)
26. Schwenk, H.: Using boosting to improve a hybrid hmm/neural network speech recognizer. In: 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings, ICASSP, vol. 2, pp. 1009–1012. IEEE (1999)
27. Shah, N.B., Wainwright, M.J.: Simple, robust and optimal ranking from pairwise comparisons. J. Mach. Learn. Res. **18**(199), 1–38 (2018)
28. Siohan, O., Ramabhadran, B., Kingsbury, B.: Constructing ensembles of asr systems using randomized decision trees. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2005, vol. 1, pp. I–197. IEEE (2005)

29. Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N., Asadpour, M.: Boosting methods for multi-class imbalanced data classification: an experimental review. J. Big Data **7** (2020)
30. Wang, C., Wu, Q., Weimer, M., Zhu, E.: Flaml: A fast and lightweight automl library (2021)
31. Wang, S., Yao, X.: Multiclass imbalance problems: analysis and potential solutions. IEEE Trans. Syst. Man Cybernet. Part B (Cybernet.) **42**(4) (2012)
32. Wauthier, F., Jordan, M., Jojic, N.: Efficient ranking from pairwise comparisons. In: International Conference on Machine Learning, pp. 109–117. PMLR (2013)
33. Wu, Q., Wang, C., Huang, S.: Frugal optimization for cost-related hyperparameters. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 10347–10354 (2021)
34. Yijing, L., Haixiang, G., Xiao, L., Yanan, L., Jinling, L.: Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data. Knowl.-Based Syst. **94** (2016)
35. Yuksel, K.A., Ferreira, T.C., Gunduz, A., Al-Badrashiny, M., Javadi, G.: A reference-less quality metric for automatic speech recognition via contrastive-learning of a multi-language model with self-supervision. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, Rhodes Island, Greece, 4-10 June 2023, pp. 1–5. IEEE (2023). https://doi.org/10.1109/ICASSPW59220.2023.10193003
36. Yuksel, K.A., Ferreira, T.C., Javadi, G., Al-Badrashiny, M., Gunduz, A.: Norefer: a referenceless quality metric for automatic speech recognition via semi-supervised language model fine-tuning with contrastive learning. In: Proc. INTERSPEECH 2023 pp. 466–470 (2023). https://doi.org/10.21437/Interspeech.2023-643
37. Zhang, Y., et al.: Google usm: Scaling automatic speech recognition beyond 100 languages. arXiv preprint arXiv:2303.01037 (2023)

Pre-training and Adverse Audio Samples for Data-Efficient Wake Word Detection

Manuel Torralbo, Ariane Méndez, Maia Agirre, and Arantza Del Pozo(✉)

Vicomtech Foundation, Basque Research and Technology Alliance (BRTA), Mikeletegi 57, 20009 Donostia-San Sebastián, Spain
{mtorralbo,amendez,magirre,adelpozo}@vicomtech.org

Abstract. This work investigates the impact of pre-training and the use of adverse audio samples on both the data efficiency and performance of end-to-end neural Wake Word Detection systems. Alongside intensive data augmentation, the proposed methodology involves pre-training Keyword Spotting models, followed by fine-tuning to recognize specific wake words by leveraging their foundational capabilities. The study also examines the inclusion of adverse audio samples resembling the target wake word. Experiments evaluate various state-of-the-art architectures to assess the effects of model size, amount of training data, model pre-training, and the incorporation of adverse audio samples on system performance. Results demonstrate that pre-training improves performance, with fine-tuned models consistently outperforming those trained from scratch, especially with limited data. Additionally, training with adverse samples resembling the wake word also enhances results by reducing false acceptance rates. These findings provide valuable insights for developing data-efficient Wake Word Detection systems.

Keywords: Wake word detection · Pre-training · Adverse audio samples · Data-efficient

1 Introduction

Wake words serve as unique phrases that activate dormant applications, marking users' initial interaction with voice-activated systems. These words are pivotal as they enable natural communication with devices at the edge, allowing users to address applications with a human-like name. This approach enhances user experience, fostering brand loyalty and association. Major companies like Amazon, Apple, and Google have heavily invested in developing and promoting their personalized assistants and their respective wake words, such as Alexa, Hey Siri, and OK Google, which have become widely recognized even outside active usage contexts.

Wake Word Detection is a specialized subset of the broader Key Word Spotting problem. While Key Word Spotting systems are designed to identify multiple keywords or commands within continuous audio streams, Wake Word Detection

A. Karpov and V. Delić (Eds.): SPECOM 2024, LNAI 15299, pp. 104–118, 2025.
https://doi.org/10.1007/978-3-031-77961-9_8

specifically focuses on recognizing a single activation phrase to initiate an interaction and requires efficient and privacy-conscious operation on edge devices.

The need for reliable Wake Word Detection systems arises from privacy concerns associated with constant listening by detection systems, as highlighted in [20]. Local deployment of these systems addresses these concerns by ensuring audio processing remains on the user's device, thus enhancing privacy and reducing dependence on cloud connectivity. This approach not only minimizes latency but also enhances system reliability and trustworthiness. Efficient Wake Word Detection engines are crucial to minimize false positives and negatives, making deployment on edge devices essential. However, these systems must also be space-efficient and utilize minimal computational resources to operate effectively on edge devices. Additionally, high-quality data is also crucial for the development of precise Wake Word Systems. Nevertheless, acquiring and annotating new data can be a time-consuming and resource-intensive process, which poses a bottleneck in their implementation.

In this work we study how pre-training and adverse audio samples impact the data requirements and performance of end-to-end neural Wake Word Detection systems. To do so, we propose a training methodology that not only utilizes data augmentation techniques but also incorporates the pre-training of Keyword Spotting models, followed by fine-tuning for specific wake word recognition. Additionally, we explore the impact of incorporating adverse audio samples similar to the target wake word. Through comprehensive experimentation, we evaluate the impacts of model size, training data volume, model pre-training, and the inclusion of adverse audio samples on system performance.

The rest of the paper is organized as follows: Sect. 2 provides some background on various Wake Word Detection approaches and techniques used to reduce the need for annotated data. In Sect. 3 and Sect. 4 the datasets and different model architectures used for experimentation are described, respectively. Section 5 focuses on the proposed training methodology. In Sect. 6 experimental results are presented, different settings and models are compared and some conclusions are drawn. Finally, Sect. 7 highlights the main contributions of this work and proposes tentative lines to develop in the future.

2 Background and Related Work

Early approaches to Wake Word Detection utilized Large Vocabulary Continuous Speech Recognition (LVCSR) systems [9,25]. In this method, the speech signal is decoded, generating lattices that represent likely sequences of phonetic units. The keyword is then searched within these lattices. This approach offers flexibility in handling non pre-defined wake words. However, the primary drawback of LVCSR-based systems lies in their considerable computational complexity, making them less suitable for resource-constrained applications or real-time processing scenarios [31]. A lighter alternative to LVCSR systems was also used, the keyword/filler approach [28,34]. This method employs two separate HMMs: one for modeling keyword audio segments and another for non-keyword (filler) segments.

The advent of deep learning led to the gradual replacement of this technologies with neural networks. The first steps in this transition where conducted by hybrid approaches that combined neural network acoustic models with HMMs [26,35]. Nevertheless, recent research has explored end-to-end neural network approaches [8], eliminating the need for HMMs. These mainly focus on the better performing convolutional neural networks [24,30] and recurrent neural networks [11,29].

Nowadays, Wake Word Detection systems are typically deployed on edge devices and must operate in real-time continuous audio streams. Consequently, neural network architectures need to balance good performance with minimal memory footprint, low computational cost and reduced latency. In this regard, optimizations such as quantization [14,16] and model conversion to streaming inference mode [29] have been developed to reduce computational demands.

Data plays a crucial role in neural networks, serving both to train model parameters and validate performance. Unfortunately, there is a lack of available corpora for Wake Word Detection [15,21]. When a new wake word is required, a data annotation process must be conducted to record audios for training, which can be time-consuming and resource-intensive.

To address these challenges, contrastive learning [36] offers an alternative way of training models. It enables the learning of distinctive speech representations by effectively utilizing both limited labeled wake word samples and abundant unlabeled speech data. Data augmentation [13,27] and synthetic audio generation techniques [2] can also help alleviate the amount of required annotated data.

To the best of our knowledge, prior research on wake word detection has not investigated the efficacy of model pre-training to enhance performance and reduce training data needs. Furthermore, the significance of employing challenging negative examples to improve model resilience against false positives has not been thoroughly explored. This work fills these gaps through a comprehensive experimental evaluation that examines how model pre-training and the integration of adverse audio samples impact various architectures, model scales, and training data sizes.

3 Datasets

This section describes each employed dataset, including their sources, characteristics, and preprocessing steps applied. Multiple datasets of different properties are used to conduct the proposed training methodology, with the aim of reducing the amount of wake word training samples required. We use Wake Word Detection datasets (Sect. 3.1) in order to validate our approach. Keyword Spotting samples (Sect. 3.2) are used for model pre-training, teaching the models general speech recognition competencies. Unknown human Speech (Sect. 3.3) and Background Noise (Sect. 3.4) data serve multiple purposes: i) each represents its own class in Keyword Spotting model pre-training, ii) they form part of the negative samples in Wake Word Detection training, and iii) they are also utilized in the data augmentation process.

All audio samples are used with a single channel and a sampling rate of 16kHz, resampled if needed employing `FFmpeg` [1]. In addition, all samples containing human speech were processed with `webrtcvad` [3] to trim audio without voice activity from the beginning and end of the recordings.

3.1 Wake Word Detection

Our proposed training approach is validated on the **Ok Aura dataset** [6] and on our own proprietary **Hey Nari dataset**, both recorded by mostly Spanish speaking volunteers. The composition of these datasets allows for robust training and evaluation of wake word detection systems, ensuring a diverse range of recording conditions, and presenting the models with both positive samples and challenging adverse samples that closely resemble the target wake word, referred to as adverse samples.

The Ok Aura dataset comprises 1,247 utterances collected from 80 distinct speakers. This collection includes positive wake word samples and adverse samples that are phonetically similar to the wake word. 151 utterances contain the wake word within a context, e.g., *"Ok Aura, ver encuentros y conferencias"*. To increase the number of positive samples, these utterances were manually cropped to isolate the wake word, resulting in a total of 510 positive samples. The remaining 737 utterances are adverse samples. The data was acquired through a web service with each speaker using their own personal microphone.

We employed a similar data acquisition process, utilizing our own audio annotation web service to collect and create the Hey Nari dataset. This dataset comprises 1357 utterances from 52 speakers, with 684 positive samples and 673 adverse samples. The most notable difference between the Ok Aura and Hey Nari datasets is the syntactic form of the adverse samples. In the Ok Aura dataset, the adverse samples are complete sentences, e.g., *"Con ese aura que tiene conseguirá lo que se proponga"*, while in the Hey Nari dataset, they are short combinations of words, e.g., *"Mi nariz"*.

3.2 Keyword Spotting

The **Google Speech Commands V2 dataset** [33] is widely used in speech recognition and keyword spotting tasks. It consists of 105,829 one second audio clips of 35 short words, collected from 2,618 speakers. The words include common commands like *"yes"*, *"no"*, *"up"*, *"down"*, digits from *"zero"* to *"nine"*, and other words covering multiple different phonemes. We balanced the dataset by randomly duplicating samples from underrepresented classes, ensuring an equal number of samples across all classes.

3.3 Unknown Speech

The **Common Voice dataset** [4] contains a vast amount of recordings from volunteers with diverse ages and accents, speaking in multiple languages. For

our experiments, we used a subset of this dataset as general unknown human speech, consisting of 25,000 audio samples in Basque, Spanish and English, the three most prominent languages in our region.

3.4 Background Noise

The employed background noise recordings come from both the **QUT-NOISE dataset** [12], and synthetic white and pink noise from the **Google Speech Commands V2 dataset** [33]. The QUT-NOISE dataset is a comprehensive background noise corpus designed for simulating noisy speech in various real-world environments. It contains 20 noise recording sessions of at least 30 min each, covering five distinct scenarios:

1. CAFE: Indoor and outdoor dining areas with background unintelligible chatter and kitchen noises.
2. HOME: Domestic settings including kitchen and living areas with typical household activities.
3. STREET: Urban intersections with varying levels of traffic and pedestrian activity.
4. CAR: Interior of a moving vehicle under different driving conditions.
5. REVERB: Large, echoic spaces with distinct acoustic properties.

4 Model Architectures

This section provides an overview of the neural network architectures evaluated in our experiments. Chosen for their cutting-edge performance and efficiency, we have implemented several model architectures for Keyword Spotting and assessed their performance in Wake Word Detection. Please refer to Sect. 5.2 for details on the specific preliminary feature extraction process required by each of these models, which take a spectrogram of the audio as input.

4.1 Broadcasting-Residual Network (BC-ResNet)

The BC-ResNet model architecture [17] has an initial 2D convolution followed by multiple broadcasted residual learning blocks. In each of these blocks the input, with frequency and temporal dimensions, is first passed through a frequency depthwise convolution and normalized using subspectral normalization [7]. This normalization method divides the frequency dimension into consecutive groups and normalizes them separately for a frequency aware normalization. The output is then averaged in the frequency dimension, leaving a single temporal dimension, and convolved depthwise. Lastly, the single feature is broadcasted or expanded back in the frequency dimension and the original input is added, forming the residual shortcut connection. A 2D separable convolution and a global average pooling layer in both frequency and temporal dimensions come after the broadcasted residual learning blocks, ending with a fully connected layer for classification. The base model BC-ResNet-1 can be scaled to create larger variants, denoted as BC-ResNet-τ, by multiplying the number of output channels throughout the network by a factor τ.

4.2 MatchboxNet

The MatchboxNet architecture [24], inspired by the QuartzNet model for automatic speech recognition [19], is also suitable for model scaling. Denoted as MatchboxNet-BxRxC, it consists of B residual blocks, each containing R sub-blocks with a 1D temporal separable convolution and C output channels. Additional temporal separable convolutions precede and succeed the residual blocks, lastly tailed by a temporal global average pooling layer and a fully connected layer for classification.

4.3 Multi-head Attention Recurrent Neural Network (MHAtt-RNN)

Based on the architecture proposed in [11], the MHAtt-RNN model [29] processes the incoming spectrograms through a series of 2D separable convolutions, followed by two bidirectional GRU [10] layers. Employing a multi-head attention mechanism [5,32], the central feature of the bidirectional GRU's output sequence is projected once per attention head via a dense layer. These projections serve as the query vectors for the attention mechanism. The attention scores are used to compute the weighted averages of the bidirectional GRU output. These attention-weighted representations are then processed through fully connected layers for final classification.

5 Proposed Training Methodology

This section outlines the multi-stage approach employed to train the different Wake Word Detection models. Section 5.1 discusses the data augmentation techniques employed to enhance data diversity and simulate real-world conditions. The methods for model-specific feature extraction and spectrogram augmentation are elaborated in Sect. 5.2. Section 5.3 outlines the pre-training process on the Keyword Spotting task, aimed at establishing foundational speech recognition capabilities. Lastly, Sect. 5.4 addresses the fine-tuning of pre-trained models specifically for Wake Word Detection.

5.1 Data Augmentation

Various data augmentation techniques are applied to the different training data sources. Keyword Spotting samples and Wake Word Detection positive and adverse samples undergo an augmentation process similar to [29,37]:

1. Temporal shifts between −100 and 100 milliseconds.
2. Resampling with a factor in the range of 0.85 to 1.15.
3. Random cropping or padding with zeros to the required fixed size.

4. Adding background noise with a probability of 0.8, scaled by a factor ranging from 0 to 0.1. The background noise is sampled randomly from seven different types: the five scenarios (CAFE, HOME, STREET, CAR, REVERB) of the QUT-NOISE dataset, synthetic audio from the Google Speech Commands V2 dataset to simulate static noise, and unknown speech from the Common Voice subset to resemble background conversations.

Since there is an abundance of unknown speech samples, they are randomly selected without the need for extensive data augmentation. However, to ensure they have similar acoustic properties to the other samples with human voice activity, background noise is added after they are randomly cropped to the desired size, as described in augmentation step 4.

Samples containing only background noise are randomly selected from six different types (those described in step 4, excluding unknown speech), randomly cropped to the desired size, and scaled by a factor ranging from 0 to 1.

5.2 Feature Extraction and Spectrogram Augmentation

The feature extraction process and specific spectrogram properties follow the recommendations provided by the respective authors for each model architecture and its scaled variants. The produced spectrograms are augmented in all cases using SpecAugment [27], applying frequency and temporal band masking, but omitting time warping:

- BC-ResNet [17]: Log Mel spectrogram with 40 Mel filterbanks with 30ms window size and 20ms overlap. BC-ResNet-1 does not use SpecAugment, but larger models BC-ResNet-{3, 6} use 2 frequency mask bands of size in the range of 0 to {5, 7}, and 2 temporal mask bands of size in the range of 0 to 20.
- MatchboxNet [24]: Mel-frequency cepstrum coefficients retaining all 64 of the computed coefficients with 25ms window size and 15ms overlap. SpecAugment is applied using 2 frequency mask bands of size in the range of 0 to 15, and 2 temporal mask bands of size in the range of 0 to 25.
- MHAtt-RNN [29]: Mel-frequency cepstrum coefficients retaining the first 40 of the 80 computed coefficients with 30ms window size and 20ms overlap. SpecAugment is applied using 2 frequency mask bands of size in the range of 0 to 7, and 2 temporal mask bands of size in the range of 0 to 25.

5.3 Model Pre-training

Inspired by [24], we pre-train the models on a Keyword Spotting task to recognize all 35 commands of the Google Speech Commands V2 dataset [33], plus two additional classes: unknown speech and background noise. Our aim is to train the models on a general and complex task as to be able to i) distinguish between a large number of keywords covering a wide range of phonemes, ii) comprehend that there is human speech that does not correspond to any of the keywords,

and iii) discern between voice activity and background noise. This knowledge is valuable, and as we demonstrate could be transferred to downstream tasks such as Wake Word Detection to improve performance.

Regarding the technical aspects, all available audio samples are used for training, with the 37 classes equally represented at training time to ensure balanced learning.

5.4 Model Fine-Tuning

The final classification layer of the pre-trained models is replaced by a single fully connected neuron with a sigmoid non-linear activation, oriented for binary classification. The models are then fine-tuned for Wake Word Detection using the following training data proportions: 40% positive samples and 60% negative samples, consisting of 24% adverse samples, 24% unknown speech, and 12% background noise. If there are insufficient positive or adverse samples to maintain these proportions, random oversampling is applied to either sample type.

The class balance is slightly skewed towards the negative samples. As expected, the main difficulty the models face is discerning between positive and adverse samples. This will be later confirmed in the experimental evaluation, where we observe high false acceptance rates, indicating that a significant number of adverse examples are misclassified as positive. To address these issues, we slightly reduced the representation of positive samples and increased the representation of adverse samples.

6 Experimental Results

All model training processes were conducted for 200 epochs using cross entropy loss and Adam optimizer [18] with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay of 10^{-2} [23]. The learning rate is linearly warmed up over the first 5 epochs, starting from 0 up to 10^{-3}, except for model fine-tuning, in which case the maximum learning rate is halved to 5×10^{-4}. Over the remaining epochs, the learning rate decays to 0 using cosine annealing [22]. Keyword Spotting model pre-training is done with a batch size of 128 and 1 s audios. Wake Word Detection models on the other hand, whether they are being fine-tuned or trained from scratch, use a batch size of 32 and 1.5 s audios.

To improve the robustness of our evaluation, we employ 10-fold cross validation on the Wake Word Detection datasets, where each fold serves once as the test data while the remaining nine are used for training. The folds are grouped by speaker, preventing audio samples recorded by the same person from appearing in both the train and test data, and are stratified by positive and adverse samples, preserving the proportion of samples for each class across folds. At each of the 10 runs, if the audio samples of the corresponding test fold are shorter than the mentioned 1.5 s, they are padded evenly with zeros. Otherwise, to emulate a real-world scenario, the samples are split in overlapping windows of 1.5 s with 100ms shifts.

Given that adverse samples are in general significantly longer than positive samples, this process results in highly unbalanced test data skewed towards the negative class. Therefore, we consider the F1 score a suitable metric to estimate overall model performance. False acceptance and rejection rates (FAR and FRR), commonly used in identification systems, are also employed to independently show where most errors occur. Note that the described evaluation method represents the worst case scenario, since all negative samples are adverse and contain speech similar to the target wake word.

Table 1. Baseline model F1 score, FAR and FRR on OK Aura and Hey Nari datasets. For reference, the top-1 test Accuracy (Acc) averaged over 5 runs on the Google Speech Commands V2 dataset (GSC V2) with 35 labels is also provided, trained as in [29]. The #Params column shows the number of parameters of each model and the #Ops column the number of multiplications and additions performed during the inference of a single sample.

Model	#Params	#Ops	OK Aura			Hey Nari			GSC V2
			F1↑	FAR(%)↓	FRR(%)↓	F1↑	FAR(%)↓	FRR(%)↓	Acc(%)↑
BC-ResNet-1	8,869	3.71M	0.883	0.110	1.086	0.979	0.393	0.218	95.020
BC-ResNet-3	53,101	21.72M	0.947	**0.077**	0.404	0.983	0.354	0.348	97.267
BC-ResNet-6	185,689	75.19M	0.957	0.116	0.269	**0.986**	0.235	**0.163**	**97.746**
MatchboxNet-3x1x64	73,473	10.84M	0.944	0.095	0.432	0.974	0.481	0.271	96.765
MatchboxNet-3x2x64	89,025	13.14M	0.950	0.125	0.335	0.985	0.232	0.227	97.047
MatchboxNet-6x2x64	135,105	19.92M	0.958	0.101	0.280	**0.986**	0.202	0.192	97.256
MHAtt-RNN	755,963	65.41M	**0.972**	0.071	**0.181**	0.981	**0.145**	0.384	97.183

6.1 Model Baselines

In Table 1 we show the baselines resulting from training the models with the proposed methodology. Even in the adverse testing environment, the models achieve excellent performance as Wake Word Detection systems. More specifically, considering the results, we provide the following observations:

- As expected, models with more trainable parameters and higher computational demands, in terms of multiplications and additions during inference, achieve better performance, especially when compared within the same model architecture.
- BC-ResNet-3 and MatchboxNet-3x2x64 models reach close to top performance while maintaining relatively low computational requirements. These models can be converted to streaming inference mode [29], making them strong candidates for real production environments on the edge.
- MHAtt-RNN, the largest tested model with approximately four times more parameters than the next largest, achieves the best performance on the OK Aura dataset but shows lower performance on the Hey Nari dataset. Its

bidirectional recurrent layers, which necessitate processing the entire input sequence, render the model incompatible with streaming inference [29]. These constraints pose challenges for deploying it in edge computing environments.
– All models exhibit markedly improved performance on the Hey Nari dataset compared to the Ok Aura dataset. We attribute this discrepancy to two primary factors: i) Despite fixed data proportions, models trained on the Ok Aura dataset encounter fewer adverse samples during training. These adverse samples consist of recordings of full sentences, which reduces the likelihood of randomly cropping to the precise adverse section. ii) *"Ok Aura"* represents a more challenging wake word due to its additional syllable.

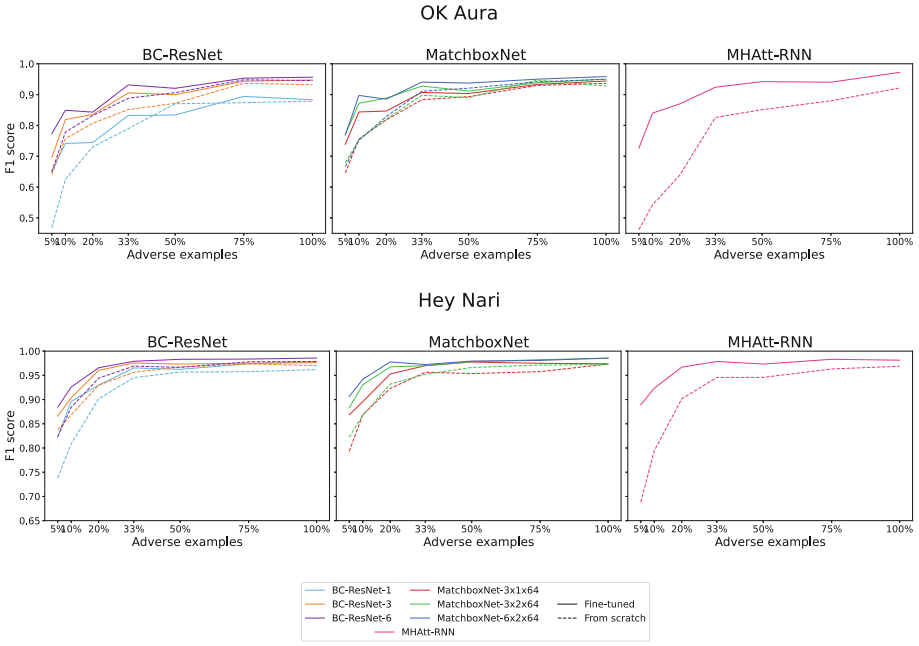


Fig. 1. Fine-tuned models and models trained from scratch using incremental proportions of the total available training data. Note that the results of the Ok Aura dataset (top row) and the Hey Nari dataset (bottom row) have different vertical scaling for better visualization.

6.2 Data Requirements and Model Pre-training

Next, we assess the data requirements for training Wake Word Detection models using our proposed approach and evaluate the impact of model pre-training. To this end, we train the models only using incremental proportions of the

total available data (5%, 10%, 20%, 33%, 50%, 75%, and 100%), comparing performance between models trained from scratch and those fine-tuned from pre-trained models. Undersampling is done randomly and without replacement, independently sampling positive and adverse samples. Models trained from scratch follow the same training procedure as fine-tuned models, but use a higher learning rate of 10^{-3}, double that of fine-tuned models, and start with randomly initialized weights.

The results, depicted in Fig. 1, indicate that around 150 to 200 positive and adverse samples (~33% of the training data in this case) are sufficient to achieve near-optimal performance. Nevertheless, as is usually the case with neural networks, adding more data beyond this continues to yield small performance improvements. Note that by using 100% of data the results shown in Table 1 are obtained.

Fine-tuned models consistently outperform models trained from scratch. This gap in performance is significant with smaller amounts of data and narrows as the available training data increases, though fine-tuned models still maintain an advantage when all data is utilized. We argue that since all trainable parameters of the pre-trained models are being adjusted during fine-tuning, performance of both training methods would likely converge if the amount of data continues to increase.

Table 2. F1 score, False Acceptance Rate (FAR) and False Rejection Rate (FRR) of fine-tuned models trained with the same amount of data, with or without adverse samples. "WW" denotes the number of Wake Words and "Adv" the number of Adverse samples used for training. The best values considering both the model and the metric are highlighted in bold.

Ok Aura	F1 score ↑		FAR(%) ↓		FRR(%) ↓	
	450 WW 0 Adv	280 WW 170 Adv	450 WW 0 Adv	280 WW 170 Adv	450 WW 0 Adv	280 WW 170 Adv
BC-ResNet-1	0.451	**0.771**	11.202	**2.784**	**0.060**	0.088
BC-ResNet-3	0.478	**0.861**	10.402	**1.391**	**0.034**	0.095
BC-ResNet-6	0.509	**0.908**	9.052	**0.844**	**0.034**	0.097
MatchboxNet-3x1x64	0.509	**0.896**	8.975	**0.925**	**0.071**	0.110
MatchboxNet-3x2x64	0.527	**0.916**	8.420	**0.695**	**0.048**	0.101
MatchboxNet-6x2x64	0.493	**0.921**	9.429	**0.680**	**0.042**	0.095
MHAtt-RNN	0.490	**0.896**	9.618	**0.941**	**0.072**	0.101
Hey Nari	**F1 score ↑**		**FAR(%) ↓**		**FRR(%) ↓**	
	610 WW 0 Adv	380 WW 230 Adv	610 WW 0 Adv	380 WW 230 Adv	610 WW 0 Adv	380 WW 230 Adv
BC-ResNet-1	0.770	**0.964**	8.337	**0.640**	**0.128**	0.392
BC-ResNet-3	0.781	**0.971**	7.991	**0.513**	**0.080**	0.343
BC-ResNet-6	0.787	**0.979**	7.589	**0.357**	**0.080**	0.210
MatchboxNet-3x1x64	0.744	**0.972**	9.690	**0.555**	**0.126**	0.270
MatchboxNet-3x2x64	0.752	**0.971**	9.288	**0.663**	**0.143**	0.267
MatchboxNet-6x2x64	0.754	**0.980**	9.023	**0.284**	**0.184**	0.287
MHAtt-RNN	0.771	**0.971**	8.436	**0.785**	**0.000**	0.079

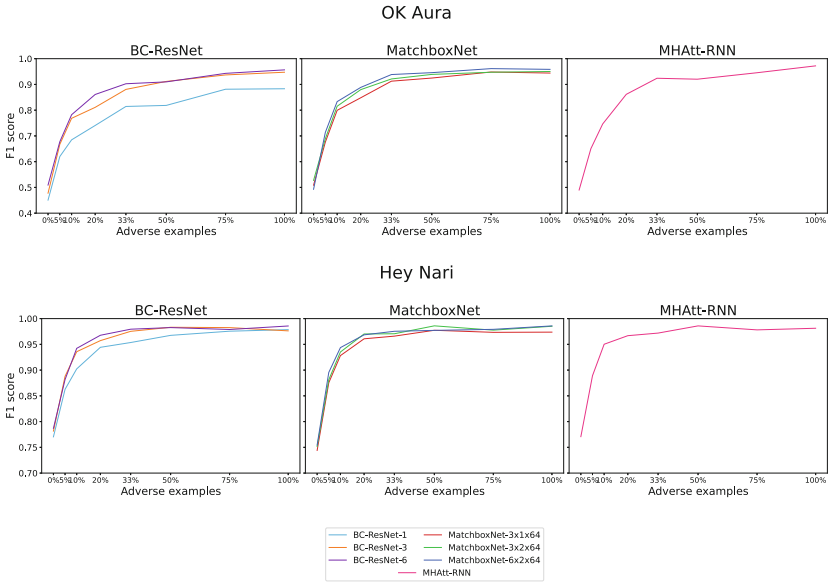


Fig. 2. Fine-tuned models trained with incremental proportions of the total adverse samples and all available positive samples of the training data. Note that the results of the Ok Aura dataset (top row) and the Hey Nari dataset (bottom row) have different vertical scaling for better visualization.

6.3 Training with Adverse Samples

To evaluate the significance of training with adverse samples, we conduct two experiments. Note that building on our previous findings, all models in these experiments are fine-tuned.

The first experiment justifies the importance of annotating adverse samples, even when it comes at the expense of recording fewer positive samples. We train the models using the same amount of annotated data, both with and without adverse samples. Models trained without adverse samples use unknown speech samples as a replacement and the maximum available wake word samples. Models trained with adverse samples use fewer positive samples and follow the data proportions established in Sect. 5.4.

Table 2 shows the results of this comparison, where we can observe that training with adverse samples undoubtedly has a positive impact on model performance. The improvement is particularly noticeable in the false acceptance rate, where models on average go from incorrectly classifying 8.52% of predictions as false positives in the Hey Nari dataset, to only 0.57%. We conclude that adverse samples force the models to learn the intrinsic and particular phonetic details of the target wake word. This comes with a trade-off, as models classify more strictly, the false rejection rate increases slightly. Note that since the models in

this comparison are trained with smaller subsets of the available data the results are not directly comparable to those presented in Table 1.

In our final experiment, we explore the amount of adverse samples required for training. To isolate their impact on model performance, we fix the number of positive samples at maximum available while using incremental proportions of adverse samples. The results, plotted in Fig. 2, confirm our previous findings that training with adverse samples positively impacts model performance. While increasing the amount of adverse data enhances model performance, diminishing returns begin to occur when using approximately 150 to 200 adverse samples (a third of the available).

7 Conclusion and Future Work

This work has investigated how pre-training and the inclusion of adverse audio samples affects the performance and data demands of end-to-end neural Wake Word Detection systems. The methodology explored has involved intensive data augmentation and initial pre-training of Keyword Spotting models, followed by fine-tuning for specific wake word recognition. Through extensive experimentation, diverse state-of-the-art architectures for Wake Word Detection have been assessed, examining the impacts of model size, training data volume, model pre-training, and the inclusion of adverse audio samples on overall performance.

The study has yielded several important insights. The impact of pre-training on Keyword Spotting has proved to be positive, with fine-tuned models consistently outperforming those trained from scratch, particularly when working with limited data. We also observed that training with adverse samples resembling the wake word notably enhances model performance, especially in reducing false acceptance rates. These samples force the models to learn nuanced phonetic characteristics of the target wake word, resulting in more robust and stringent models, albeit with a slight increase in false rejection rates. The paper provides extensive analysis and insights on the training methodologies, architectures, model sizes, and data prerequisites necessary for developing efficient and performing Wake Word Detection systems suitable for on edge deployment.

Moving forward, our goals include continuing to enhance model performance while minimizing training data needs. Key objectives involve exploring the generation of synthetic wake word samples and automating the collection of adverse samples, thereby eliminating the need for manual annotation.

Acknowledgments. This work has received funding from the Department of Economic Development and Infrastructure of the Basque Government under grant number KK-2022/00102 (BERREKIN).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. FFmpeg. https://ffmpeg.org/. Accessed 10 Jul 2024
2. Piper. https://github.com/rhasspy/piper. Accessed 15 Jul 2024
3. webrtcvad. https://pypi.org/project/webrtcvad/. Accessed 10 Jul 2024
4. Ardila, R., et al.: Common voice: a massively-multilingual speech corpus. arXiv preprint arXiv:1912.06670 (2019)
5. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
6. Bonet, D., et al.: Speech enhancement for wake-up-word detection in voice assistants. In: Proceedings of the IberSPEECH 2021, pp. 41–45 (2021). https://doi.org/10.21437/IberSPEECH.2021-9
7. Chang, S., Park, H., Cho, J., Park, H., Yun, S., Hwang, K.: Subspectral normalization for neural audio data processing. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 850–854. IEEE (2021)
8. Chen, G., Parada, C., Heigold, G.: Small-footprint keyword spotting using deep neural networks. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4087–4091 (2014). https://doi.org/10.1109/ICASSP.2014.6854370
9. Chen, G., Yilmaz, O., Trmal, J., Povey, D., Khudanpur, S.: Using proxies for OOV keywords in the keyword search task. In: 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 416–421. IEEE (2013)
10. Cho, K., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
11. De Andrade, D.C., Leo, S., Viana, M.L.D.S., Bernkopf, C.: A neural attention model for speech command recognition. arXiv preprint arXiv:1808.08929 (2018)
12. Dean, D., Kanagasundaram, A., Ghaemmaghami, H., Rahman, M.H., Sridharan, S.: The qut-noise-sre protocol for the evaluation of noisy speaker recognition. In: Proceedings of the 16th Annual Conference of the International Speech Communication Association, Interspeech 2015, pp. 3456–3460. International Speech Communication Association (2015)
13. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017)
14. Gong, Y., Liu, L., Yang, M., Bourdev, L.: Compressing deep convolutional networks using vector quantization. arXiv preprint arXiv:1412.6115 (2014)
15. Hossain, D., Sato, Y.: Efficient corpus design for wake-word detection. In: 2021 IEEE Spoken Language Technology Workshop (SLT), pp. 1094–1100. IEEE (2021)
16. Jacob, B., et al.: Quantization and training of neural networks for efficient integer-arithmetic-only inference. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2704–2713 (2018)
17. Kim, B., Chang, S., Lee, J., Sung, D.: Broadcasted residual learning for efficient keyword spotting. arXiv preprint arXiv:2106.04140 (2021)
18. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
19. Kriman, S., et al.: QuartzNet: deep automatic speech recognition with 1d time-channel separable convolutions. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6124–6128. IEEE (2020)

20. Lau, J., Zimmerman, B., Schaub, F.: Alexa, are you listening? privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. Proc. ACM Hum.-Comput. Interact. **2**(CSCW), 1–31 (2018)
21. López-Espejo, I., Tan, Z.H., Hansen, J.H., Jensen, J.: Deep spoken keyword spotting: an overview. IEEE Access **10**, 4169–4199 (2021)
22. Loshchilov, I., Hutter, F.: SGDR: stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
23. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
24. Majumdar, S., Ginsburg, B.: Matchboxnet: 1d time-channel separable convolutional neural network architecture for speech commands recognition. arXiv preprint arXiv:2004.08531 (2020)
25. Miller, D.R., et al.: Rapid and accurate spoken term detection. In: Interspeech. vol. 7, pp. 314–317 (2007)
26. Panchapagesan, S., et al.: Multi-task learning and weighted cross-entropy for DNN-based keyword spotting (2016)
27. Park, D.S., et al.: Specaugment: a simple data augmentation method for automatic speech recognition. arXiv preprint arXiv:1904.08779 (2019)
28. Rose, R.C., Paul, D.B.: A hidden Markov model based keyword recognition system. In: International Conference on Acoustics, Speech, and Signal Processing, pp. 129–132. IEEE (1990)
29. Rybakov, O., Kononenko, N., Subrahmanya, N., Visontai, M., Laurenzo, S.: Streaming keyword spotting on mobile devices. arXiv preprint arXiv:2005.06720 (2020)
30. Sainath, T.N., Parada, C.: Convolutional neural networks for small-footprint keyword spotting. In: Interspeech, pp. 1478–1482 (2015)
31. Shan, C., Zhang, J., Wang, Y., Xie, L.: Attention-based end-to-end models for small-footprint keyword spotting. arXiv preprint arXiv:1803.10916 (2018)
32. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
33. Warden, P.: Speech commands: a dataset for limited-vocabulary speech recognition. arXiv preprint arXiv:1804.03209 (2018)
34. Wilpon, J.G., Miller, L.G., Modi, P.: Improvements and applications for key word recognition using hidden markov modeling techniques. In: [Proceedings] ICASSP 91: 1991 International Conference on Acoustics, Speech, and Signal Processing, pp. 309–312. IEEE (1991)
35. Wu, M., et al.: Monophone-based background modeling for two-stage on-device wake word detection. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5494–5498. IEEE (2018)
36. Xi, Y., Yang, B., Li, H., Guo, J., Yu, K.: Contrastive learning with audio discrimination for customizable keyword spotting in continuous speech. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 11666–11670. IEEE (2024)
37. Zhang, Y., Suda, N., Lai, L., Chandra, V.: Hello edge: keyword spotting on microcontrollers. arXiv preprint arXiv:1711.07128 (2017)

Cross-Lingual Summarization of Speech-to-Speech Translation: A Baseline

Pranav Karande(✉), Balaram Sarkar, and Chandresh Kumar Maurya

Indian Institute of Technology Indore, Indore, India
pranav.3943@gmail.com

Abstract. Cross-lingual speech-to-speech translation, which enables spoken language conversion from one language to another, plays a pivotal role in overcoming language barriers and promoting cross-cultural communication. The proliferation of multimedia content poses challenges for audiences to efficiently consume extended audio, such as news broadcasts, academic lectures, and political speeches. To address this, we propose a novel investigation of summarization in the context of cross-lingual speech-to-speech translation (S2S-Summ) with a focus on low-resource Indic languages. To the best of our knowledge, this task has not been explored in prior research. We develop and present a semi-synthetic dataset of translated summaries in Hindi (Hi), Bengali (Bn), Gujarati (Gu), and Tamil (Ta) languages, alongside baseline models for this task. The performance of our models is evaluated using metrics such as BERTScore, ROUGE, and UniEval. Our study aims to catalyze further exploration in this area, facilitating streamlined access to multilingual audio content and enhancing information dissemination across linguistic boundaries. Code and data is available at https://github.com/pranavkarande/S2S-Summ.

Keywords: Speech summarization · Low-resource languages

1 Introduction

In our interconnected world, transcending linguistic and cultural boundaries in communication is essential. Speech, the most natural form of human interaction, is key to this global exchange of ideas [1]. We envision a world where conversations flow effortlessly across borders, cultural nuances are seamlessly conveyed, and information is easily accessible to all, regardless of language. However, significant challenges remain. Language barriers hinder understanding, and the vast amount of information from news, social media, academic lectures, and business presentations can be overwhelming.

The field of Natural Language Processing (NLP) has experienced remarkable advancements, paving the way for innovative speech technologies that bridge

P. Karande B. Sarkar — contributed equally

A. Karpov and V. Delić (Eds.): SPECOM 2024, LNAI 15299, pp. 119–133, 2025.
https://doi.org/10.1007/978-3-031-77961-9_9

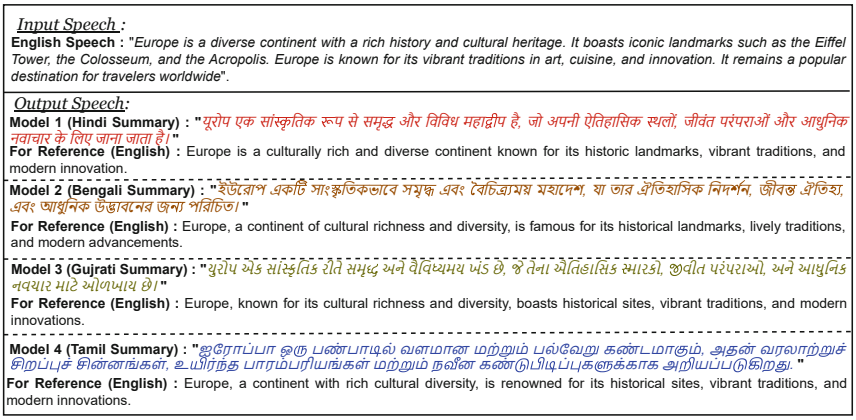


Fig. 1. Task diagram showing sample input speech in English and output speech summary in target languages Hindi, Bengali, Gujarati, and Tamil.

the language divide. Speech-to-Speech Translation (S2ST) systems [2–4] have emerged as a powerful tool for real-time communication across languages. By directly translating spoken utterances, S2ST systems have revolutionized interactions in diverse settings, from international conferences to business meetings.

Speech-to-Text Translation (ST), another valuable approach, directly translates spoken language into text in another language. While traditionally implemented as a two-step process, advancements in end-to-end ST models now enable the direct conversion of speech into text in another language, bypassing the intermediate text representation stage. This approach offers increased efficiency and can potentially improve translation accuracy [5,6]. Both S2ST and ST may result in lengthy translations, often overwhelming the listeners. Summarization is crucial in these situations, distilling the essence of the speech and reducing information overload. It enables listeners to quickly grasp key points and make informed decisions, facilitating efficient communication and improving comprehension.

To address these limitations and enhance communication, we propose a novel approach called cross-lingual speech-to-speech summarization (S2S-Summ) (Fig. 1). S2S-Summ goes beyond direct translation, generating concise and informative summaries in a different target language. By capturing the key points, arguments, and overall meaning of the original speech, S2S-Summ mitigates information overload and unnatural phrasing found in traditional translation methods. Our research pioneers the exploration of this task. Our main contributions are:

a. To propose baseline cascaded models that utilize automatic speech recognition (ASR), machine translation (MT), summarization (Summ), and text-to-speech (TTS) for this novel task.
b. A semi-synthetic dataset for the S2S-Summ task, with an emphasis on low-resource Indic languages namely Hindi (Hi), Bengali (Bn), Gujarati (Gu), and Tamil (Ta).

c. To perform automatic and human evaluations on the generated summaries from the proposed methods for a comprehensive evaluation.

2 Problem Definition

Speech-to-Speech Translation Summarization (S2S-Summ) is the process of summarizing a paragraph from a source language speech (S) to a target language speech (Z_t). The model for this task is trained to optimize the log-likelihood of the reference summary based on the input speech. Formally, it is defined as follows: Given the dataset $D = \{(x, y)\}_{i=1}^{n}$, where $x = \{x_1, x_2, .., x_s\}$ is the input speech feature vector and $y = \{y_1, y_2, .., y_t\}$ is the output speech feature vector. The model is optimized to minimize the negative log-likelihood $-\log p(y|x; \theta)$ where the conditional probability is defined as:

$$p(y|x; \theta) = \prod_{k=1}^{n} p(y_k|y_{<k}, x; \theta) \tag{1}$$

In the above equation, θ denotes the model parameters. A sequence-to-sequence (seq2seq) model can address the problem in (1), but direct application of the model presents challenges. These include (a) the need for the model to learn how to align speech and text across different languages, and (b) the requirement to produce an effective and ideally abstractive summary. **Abstractive summarization** involves the generation of new sentences to convey the main ideas whereas **extractive summarization** involves selecting important sentences or passages directly from the source text.

3 Related Works

Cross-lingual S2S-Summ remains largely unexplored in natural language processing tasks. Recent progress in text-to-text and speech-to-text summarization offers some insights, but challenges like multilingual accuracy and cultural nuances persist. This section presents works on (1) text-to-text summarization and (2) speech-to-text summarization.

3.1 Text-to-Text Summarization

Recent advancements in summarization encompass a diverse array of methodologies. These approaches typically instantiate their encoder-decoder framework by selecting from options such as RNN [7], Transformer [8,9], or GNN [10] as encoders, and either non-auto-regressive [11,12] or auto-regressive decoders [13,14]. Despite their efficacy, these models predominantly operate at the sentence level, employing individual scoring processes that favor the highest-scoring sentence, which may not necessarily be the optimal choice for forming a summary. [15] proposed a graph-based abstractive summarization method for biomedical text.

State-of-the-art methods often rely on leveraging pre-trained large sequence-to-sequence (Seq2seq) language models such as BART and T5. These models are

then fine-tuned using datasets specific to abstractive or extractive summarization. However, [16] introduced a novel method that utilizes encoder-only language models like RoBERTa [17], elevating them to decoder modules. Subsequently, these encoder-decoder models can be fine-tuned for various downstream tasks.

3.2 Speech-to-Text Summarization

Only a few studies have delved into speech-to-text summarization tasks previously. [18] employs a restricted self-attention mechanism to facilitate the processing of lengthy input audios within a transformer architecture. Initially, the authors train a randomly initialized model for Automatic Speech Recognition (ASR), followed by training it for S2T abstractive summarization. Similarly, [19] capitalizes on a Text-to-Text (T2T) abstractive summarization corpus, incorporating a text-to-speech voice synthesizer for data augmentation. Both studies yield superior outcomes compared to robust cascade baselines. In the work of [20], the decoder is transferred from a T2T summarizer to the S2T model to introduce a cross-modal adapter that aligns speech and textual features. Following pre-training of the adapter, they fine-tune the entire S2T summarizer on the BNews corpus and select the best-performing checkpoint. This approach represents a significant step forward in S2T abstractive summarization, offering a promising avenue for improving summarization performance by effectively integrating speech and text representations.

4 Dataset Synthesis

Given that the task of summarizing speech from one language to speech in another language is a novel challenge, there is a lack of existing datasets designed specifically for this task. Most available datasets focus on either speech translation or text summarization separately. A study on publicly available datasets on Kaggle reveals that out of over 1581 datasets related to natural language processing, the number of datasets that address MT is 729, audio classification is 325, ASR is 488, S2T is 28, Text summarization in the same language is 11, and speech to text summarization in a different language is 0.

Hence, we curate a semi-synthetic dataset for the task of summarizing speech-to-speech translation from MuST-C [21] dataset of English to German language ST task, which is created from TED,[1] talks. As we aim to summarize En speech to Hi, Bn, Gu and Ta speech, the transcript of the former language is summarized with a pre-trained summarizer BART [22],[2] trained on CNN/Daily Mail [23] dataset to generate En summaries, which is then translated into text of the target language using En to Hi, Bn, Gu and Ta translator, which use Google translation API[3]. Further by the use of Indic-TTS [24][4] from AI4Bharat we synthesize text summaries in Hi, Bn, Gu and Ta into speech summaries.

[1] https://www.ted.com/.
[2] https://huggingface.co/facebook/bart-large-cnn.
[3] https://doctranslator.com/.
[4] https://github.com/AI4Bharat/Indic-TTS.

Table 1. Statistics of the synthetic dataset for S2TSumm task for $En \rightarrow Hi, Bn, Gu, Ta$ language pairs. T, X_t, V & Y_t are explained in §4.1.

En →	Split	Speech (Hrs)	Para (in K)	Tokens (in K)			
				T	X_t	V	Y_t
Hi	Train	360	13.5	4663	5001	647	838
	Test	49	1.6	502	598	80	95
Bn	Train	360	13.5	4663	5723	647	1003
	Test	49	1.6	502	673	80	102
Gu	Train	360	13.5	4663	4782	647	795
	Test	49	1.6	502	574	80	89
Ta	Train	360	13.5	4663	6154	647	1147
	Test	49	1.6	502	724	80	113

As the summaries and their speech on *only the target side* are synthetically generated, we call it a *semi-synthetic* dataset. All the data created using the above-mentioned models is manually validated. A total of seven annotators in the age group of 18–28 are employed for validation among which 5 are male and 2 are female, all proficient in respective indic languages. As compared to other models such as mT5 [25], we use BART to synthesize data since the summaries from BART surpass the others in clarity and depth upon manual validation.

4.1 Dataset Statistics

The synthetic dataset created for S2S-Summ contains the speech (S) and its transcripts (T) in En language, with its translated text (X_t) in target language (where $t = h$ means Hi, b means Bn, g means Gu and t means Ta) along with its summary in En text (V), target language text (Y_t) and target language

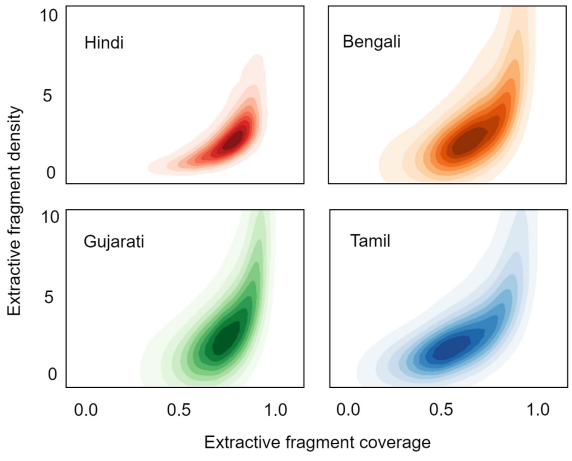


Fig. 2. Density estimate of extractive diversity scores as described in Sect. 3.2 using kernel density estimation on $En \rightarrow Hi$, Bn, Gu, and Ta summaries.

speech (Z_t). The summary in target language text (Y_t) and speech (Z_t) acts as a reference summary for the text and speech summary generated from the models. The detailed statistics are shown in Table 1.

4.2 Data Diversity

To extract the nature of a dataset, [26] define three measures, which we use here for the study. In the multi-document setting, we combine the source documents into one input by joining them together. Extractive fragment coverage (EFC) is the measure of how many words in the summary come directly from the source material, indicating the summary's reliance on the original text:

$$EFC(T, Y) = \frac{1}{|Y|} \sum_{f \in F(T,Y)} |f| \tag{2}$$

In the above eq. T is raw text, Y is summary, and F(T, Y) is the set of all token sequences that are identified as extractive. In a greedy approach, the process marks a sequence of source tokens as extractive if it serves as a prefix for the rest of the summary. Similarly, density (δ) measures the average length of the extractive fragment to which each word in the summary belongs to:

$$\delta = \frac{1}{|Y|} \sum_{f \in F(T,Y)} |f|^2 \tag{3}$$

These numbers are plotted using kernel density estimation in Fig. 2 for all pairs of languages. A large variation on the y-axis indicates differences in the average length of source sequences included in the summary. Meanwhile, variations along the x-axis reflect the average length of extractive fragments associated with words in the summary. Regarding the y-axis (fragment density), our dataset exhibits fluctuations in the average length of copied sequences, indicating diverse styles of word sequence arrangement.

5 Methodology

We propose four different cascaded approaches for S2S-Summ as shown in Fig. 3. All these approaches take *speech* as input in *language s* and output *speech summary* in *language t* as shown in Fig. 3. In all of the methods, the end component used is a TTS model, hence we have different approaches based on the intermediate components for each setting. All the proposed models are optimized to minimize the negative log-likelihood for the reference translated summary.

SS-Summ-Trans: Speech-to-Speech translated summarization (SS-Summ-Trans) is the approach of translating the textual summary of input speech in language s to the target language t. This model can be realized by transcribing T the speech S of language s using a pre-trained ASR model, then summarizing it as V in the same language using a pre-trained Summ model, further

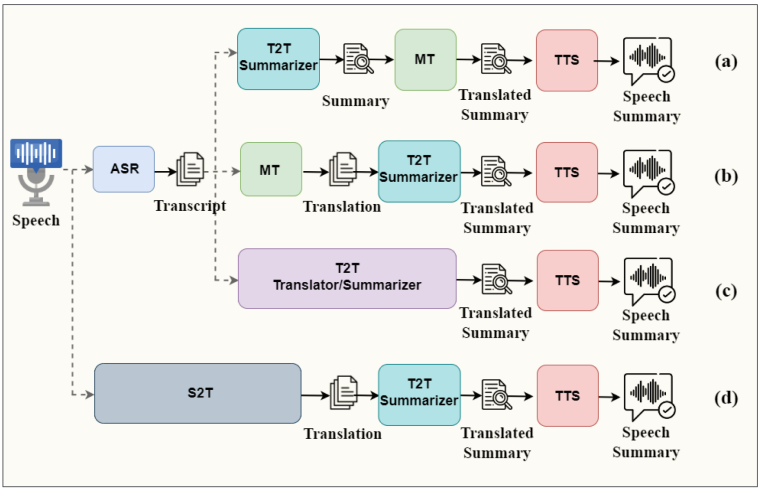


Fig. 3. Proposed methods for summarizing Speech-to-Speech Translation: (a) SS-Summ-Trans, (b) SS-Trans-Summ, (c) S2STrans/Summ, and (d) S2T-S-Summ.

translating the summary Y_t into the target language t using a pre-trained MT model, and then finally synthesizing the translated summary into speech Z_t using a pre-trained TTS model. The conditional probability distribution of the SS-Summ-Trans model is as follows:

$$p(Z_t|S) = p(T|S)p(V|T,S)p(Y_t|V,T,S)p(Z_t|Y_t,V,T,S) \tag{4}$$

SS-Trans-Summ: Speech-to-Speech summarized translation (SS-Trans-Summ) summarizes the speech in language s to speech in language t by converting speech into the transcript using pre-trained ASR models, followed by translating the transcript text of language s into language t using pre-trained MT, then summarizing the text of language t by fine-tuning Summ model on $(X_t \rightarrow Y_t)$, and then synthesizing the translated summary into speech Z_t with a pre-trained TTS model. SS-Trans-Summ model factors the conditional probability to include transcripts and translations in (1) as:

$$p(Z_t|S) = p(T|S)p(V|T,S)p(X_t|V,T,S)p(Z_t|X_t,V,T,S) \tag{5}$$

S2STrans/Summ: The Speech-to-Speech Translation/Summarization method (S2STrans/Summ) concurrently summarizes and translates into the target language t from transcriptions of input speech in language s by fine-tuning the Summ model on our dataset. **S2T-S-Summ:** Summarization of direct end-to-end speech-to-text translation (S2T-S-Summ) approach summarizes the text in the target language t which is translated from input speech in language s with the help of direct ST models.

6 Experiments

This section presents the details of (1) the models and hyperparameters we follow for our task, (2) the training configurations, and (4) the metrics used.

6.1 Models and Hyperparameters

All the methods described in Sect. 5 have Indic-TTS [24] by AI4-Bharat as the last component to synthesize the speech in the target language. The detailed training configurations of all the intermediate models used in all the approaches are as follows:

SS-Summ-Trans trains the model using *English* speech, transcripts, and summaries, along with translated target language summaries. For ASR, we train **Transfomer** [27] from *Fairseq toolkit* [28] on ($S \rightarrow T$), which has an input embedding dimension of 256, 12 encoder layers, 6 decoder layers, a hidden dimension of 2048 for feedforward sub-layers, 4 attention heads, and utilizes the ReLu activation function. For the Summ model on target language transcripts, we employ **mBART** finetuned on ($X_t \rightarrow Y_t$) and **mT5** many-to-many model pre-trained on the target languages. For MT, we explore two pre-trained models and one finetuned on our data ($T \rightarrow X_t$): **mBART** [29] which is fine-tuned on our dataset utilizes a standard Seq2Seq Transformer, **madlad-400** [30] architecture uses a Sentence Piece Model with 256k tokens shared between the 32-layer encoder and decoder each, where all input sentences begin with a <2xx> token to denote the target language and **seamless-m4T** [31] employs the transformer encoder-decoder model from NLLB [32], featuring a model dimension of 1024, FFN dimension of 8192, 16 attention heads, and 24 layers for both encoder and decoder.

SS-Trans-Summ trains the model on a quadruple of *En* speech, *En* transcripts, target language transcripts, and target language summary. To realize this cascaded approach, we employ ASR, MT, and Summ models. We utilize the same pre-trained *Transformer* model for *ASR* as in SS-Summ-Trans. The translation of *En* summaries to the target language summaries uses the same *MT* models employed in *SS-Summ-Trans* and the *summarization* model employs: **BART** and **T5** [33] both of which are pre-trained and finetuned on ($T \rightarrow V$).

S2STrans/Summ trains the model on a triplet of *En* speech, its transcript, and the target language summary. Again, a pre-trained *Transformer* is used for training the *ASR* like above. We then *finetune* the pre-trained *mBART* model for jointly training *En* transcript to synthesize the summary that translates and summarizes together into the target language ($T \rightarrow Y_t$).

S2T-S-Summ trains the model on a triplet of *En* speech, its target language translation, and its summary. We train the model over a pre-trained Transformer for S2T ($S \rightarrow X_t$) as used above for ASR, and finetune the pre-trained mBART model for summarizing into the target language ($X_t \rightarrow Y_t$).

6.2 Training Configuration

All the data processing for the generation of synthetic dataset and model training for all the proposed methods and settings is done on the following machines: NVIDIA GeForce RTX-A5000 GPU with 24 GB of VRAM and RTX-A4500 with 20GB of VRAM, respectively.

6.3 Metrics

We employ various metrics for evaluating the results of the proposed approaches. To evaluate intermediate ASR results, we use *WER* [34], and to evaluate MT results, we get *BLEU* score using sacrebleu [35]. For evaluating Summarization results, we use varied metrics: firstly *ROGUE*,[5] [36] generates Rouge1 and Rouge2 based on n-gram, RougeL on word sequences, and RougeLSum considers new lines as well, second is *F1-Score* from *BERTScore*,[6] [37] which matches the reference and system-generated summary using cosine similarity and the third is *UniEval*,[7] [38] which evaluates *Coherence* and *Consistency* both of which compares reference and system-generated summary, also *Fluency* checks the quality of system-generated summary, and *Relevance* tells whether the summary generated holds important information from input sentences. All metrics are in the range of 0–100. Although there are ways to evaluate the correctness of speech synthesis, in S2S-Summ we assess it through the intermediate-generated text summaries.

7 Experimental Results

Table 2 summarizes the results of all the proposed approaches for the pair $En \rightarrow Hi$. In Table 3, we project only the scores of the best-performing model and its setting for $En \rightarrow Bn$, Gu, and Ta languages. The ROUGE metric doesn't support Bn, Gu, and Ta, so we use UniEval and BERTScore (BT) to evaluate these instead. We present the metrics on MT and Summ models when compared to the ground-truth summary of the dataset. Table 2 presents metrics for all proposed approaches, but we don't compare all outcomes due to differing settings. Instead, we compare model configurations within the same setting across all metrics. Due to the limitation of the speech-to-speech summarization evaluation metric, we analyze only the text summaries for our task. In SS-Trans-Summ, S2STrans/Summ, and S2T-S-Summ, the final metric is ROGUE (R1, R2, RL, RLSum) since summarization is performed last. In SS-Summ-Trans, it is BLEU because translation is performed last. UniEval and BT are also used as final metrics for all approaches, evaluated on the predicted and reference summaries.

From Table 2 we observe that the final metrics ROUGE, BLEU, and BT in general are incongruent with the metric UniEval. That means, qualitative

[5] https://huggingface.co/spaces/evaluate-metric/rouge.
[6] https://github.com/Tiiiger/bert_score.
[7] https://github.com/maszhongming/UniEval.

Table 2. Results of all the model settings for the proposed methods for *En* Speech to *Hi* Text Summary are given below. In ST-TransSumm and ST-SummTrans, Transformer is used for ASR with 3.84 WER. Best results are underlined for all metrics and the result is in boldface for the final metric. (R1: Rouge1, R2: Rouge2, RL: RougeL, and RLSum: RougeLSum from Rogue; Coh: Coherence, Con: Consistency, Flu: Fluency, and Rel: Relativity from UniEval; and BT: F1-Score from BertScore. '-'denotes the metric is not applicable and 'blank space'follows the values above.)

Models	MT	Summ				UniEval				BT
	BLEU	R1	R2	RL	RLSum	Coh	Con	Flu	Rel	
Ground-truth	28.70	–	–	–	–	80.91	82.71	86.39	82.87	95.82
ST-Summ-Trans										
Transformer										
+ BART + mBART	12.14	52.99	37.65	44.58	51.03	74	84.23	88.41	73.36	75.17
+ BART + madlad-400	15.98					76.27	83.66	87.59	73.95	75.76
+ BART + seamless-m4T	**19.25**					78.26	80.28	86.33	72.99	**76.85**
+ T5 + mBART	6.90	35.52	18.98	28.65	33.97	75.26	**84.85**	88.86	72.80	70.85
+ T5 + madlad-400	7.54					73.88	83.48	87.39	69.80	70.69
+ T5 + seamless-m4T	9.33					77.51	79.43	86.43	70.33	71.84
Ground-truth	-	16.32	10.87	19	20.67	80.91	82.71	86.39	82.87	95.82
ST-Trans-Summ										
Transformer										
+ mBART + mBART	34.56	11.83	3.66	11.75	11.74	60.06	76.08	81.15	74.10	68.66
+ madlad-400 + mBART	37.06	**21.56**	8.07	13.5	12.48	59.86	76.45	80.06	73.92	69.07
+ seamless-m4T + mBART	51.08	12.58	**14.03**	**15.58**	**15.50**	60.09	76.40	80.91	74.01	69.65
+ mBART + mT5	34.56	5.74	1.79	5.64	5.67	79.42	81.25	86.88	50.97	66.64
+ madlad-400 + mT5	37.06	5.83	1.86	5.71	5.69	79.88	81.48	86.91	60.84	66.59
+ seamless-m4T + mT5	51.08	5.91	1.74	5.81	5.83	80.35	82.03	86.95	60.34	67.34
S2STrans/Summ										
Transformer										
+ finetuned mBART	-	12.27	4.13	12.05	12.03	**80.65**	82.34	85.93	71.54	74.82
+ finetuned mT5	-	3.13	0.60	3.08	3.07	77.12	84.05	**89.97**	**76.54**	66.56
S2T-S-Summ										
Transformer(S2T)										
+ mBART	40.41	17.216	9.602	14.05	13.93	79.63	81.50	86.32	69.15	72.62
+ mT5	40.41	5.09	1.49	5.02	5.03	79.33	81.18	87.24	74.26	68.66

and quantitative metrics disagree with each other. To understand the best-performing method according to both the scores, we plot UniEval and BT of the models performing best on BT (for each setting) on a radar graph and find the model that covers the maximum area in the bounded region for each setting as shown on Fig. 4. The scores of the best-performing model and setting are shown in Table 3. In the first setting, SS-Summ-Trans, as seen in Table 3,

BART + mBART performs best for $En \rightarrow Bn$, Gu and Ta language pairs when compared with BLEU and BT metrics. When examining the UniEval scores in setting one, we find that BART + mBART beats all other models on all language pairs.
In setting two, for Bn and Gu, SS-Trans-Summ, mBART + mT5 outperforms among all models in terms of UniEval and BT, and the model seamless-m4T + mT5 outperforms best for the language Ta. In setting three, S2STrans/Summ, mBART performs adequately well for all languages. Similarly in setting four, S2T-S-Summ, mBART performs better on the BT score and mT5 performs best in terms of UniEval scores.

If summarization is done last, SS-Trans-Summ with seamless-m4T + mBART is ideal. If translation is last, SS-Summ-Trans with BART + seamless-m4T performs best (Table 2). Using BT and UniEval metrics, ST-Summ-Trans with BART + seamless-m4T outperforms all models in BT, while S2STrans/Summ with mT5 excels in UniEval. Figure 4 shows that setting one (red line) covers the maximum area in all graphs, indicating it performs best across all language pairs.

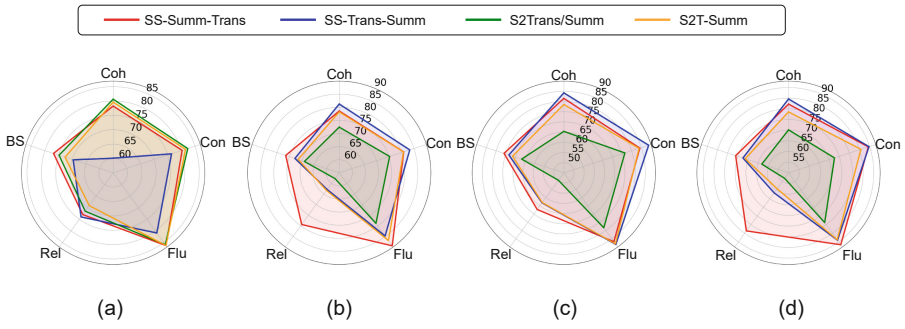


Fig. 4. UniEval and BERTScore of best-performing model from all settings for the proposed methods on En speech to (a) Hi, (b) Bn, (c) Gu and (d) Ta summaries respectively shown on radar graph (Coh: Coherence, Con: Consistency, Flu: Fluency, Rel: Relevance and BS: BERTScore).

Table 3. Results of the best performing models and setting on $En \rightarrow Bn$, Gu and Ta pairs.

Language	Models	Summ					MT		BT
		Par	R1	R2	RL	RLSum	Par	BLEU	
Bn	**SS-Summ-Trans** Transformer + BART + mBART	600mn	53	37.66	44.58	51.03	611 mn	16.23	76.53
Gu	**SS-Summ-Trans** Transformer + BART + mBART	600mn	53	37.66	44.58	51.03	611 mn	20.34	76.17
Ta	**SS-Summ-Trans** Transformer + BART + mBART	600mn	53	37.66	44.58	51.03	611 mn	14.86	77.25

8 Human Evaluation

For human evaluation, we compute adequacy and fluency scores as shown in Fig. 5. Four male evaluators, aged 18–30, proficient in English and the targeted Indic language, were employed to ensure demographic representation. The adequacy score assesses how well the summary captures the key information from the original text, while the fluency score evaluates readability and coherence in the target language. Five hundred summaries from the best-performing models for each language pair are manually evaluated. As can be seen in Fig. 5 the model performs best on Hi in terms of adequacy score and Gu in fluency score. The performance on Bn seems adequately well whereas Ta scores are comparatively less than other languages. One possible reason for this could be due to its morphologically complex nature. As Ta has many ways to represent a set of words or sentences, the summarization can have multiple possible outputs.

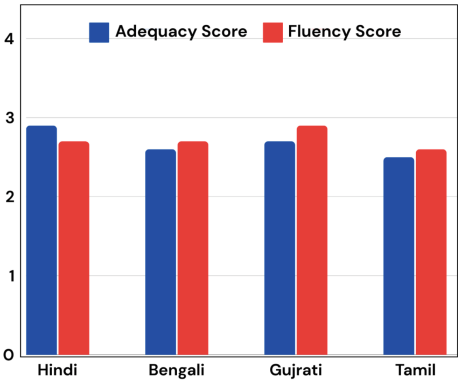


Fig. 5. Human evaluation of adequacy and fluency of the best-performing model of each language pair.

9 Conclusion

The current study introduces cascade models aimed at summarizing cross-lingual speech-to-speech translation. Our investigation reveals that this is the first instance of addressing such a task for any language pair. We present four distinct approaches to tackle this task and construct a synthetic dataset containing English speech to target speech and text summaries on four Indic languages for analysis. While envisioning the development of a dedicated dataset and further exploration of various model configurations, our future objectives include

the creation of an end-to-end S2S-Summ model. We extend an open challenge to researchers to devise an evaluation metric for this task, where both translation and summarization contribute equally to the model-generated summary. In short, the promising results of this cross-lingual multimodal task hold the potential for advancing the convergence of natural language processing and spoken language translation, paving the way for researchers to explore new avenues.

References

1. Fitch, W.T.: The evolution of language: a comparative review. Biol. Philos. **20**, 193–203 (2005)
2. Popuri, S., et al.: Enhanced Direct Speech-to-Speech Translation Using Self-supervised Pre-training and Data Augmentation (2022). Interspeech
3. Wang, Y., Bai, J., Huang, R., Li, R., Hong, Z., Zhao, Z.: Speech-to-Speech Translation with Discrete-Unit-Based Style Transfer. arXiv:abs/2309.07566 (2023)
4. Inaguma, H., et al.: UnitY: two-pass direct speech-to-speech translation with discrete units. arXiv:abs/2212.08055 (2022)
5. Zhou, G., Lam, T., Birch, A., Haddow, B.: Prosody in Cascade and Direct Speech-to-Text Translation: a case study on Korean Wh-Phrases. Findings (2024)
6. Sarkar, B., Maurya, C.K., Agrahri, A.: Direct speech to text translation: bridging the modality gap using SimSiam. In: Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023), pp. 250–255 (2023)
7. Zhou, Q., Yang, N., Wei, F., Huang, S., Zhou, M., Zhao, T.: Neural document summarization by jointly learning to score and select sentences. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 654–663 (2018)
8. Zhong, M., Liu, P., Wang, D., Qiu, X., Huang, X.: Searching for effective neural extractive summarization: what works and what's next. In: Annual Meeting of the Association for Computational Linguistics (2019)
9. Wang, D., Liu, P., Zhong, M., Fu, J., Qiu, X., Huang, X.: Exploring domain shift in extractive text summarization. arXiv:abs/1908.11664 (2019)
10. Wang, D., Liu, P., Zheng, Y., Qiu, X., Huang, X.: Heterogeneous Graph Neural Networks for Extractive Document Summarization. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 6209–6219 (2020)
11. Narayan, S., Cohen, S.B., Lapata, M.: Ranking sentences for extractive summarization with reinforcement learning. In: North American Chapter of the Association for Computational Linguistics (2018)
12. Arumae, K., Liu, F.: Reinforced extractive summarization with question-focused rewards. arXiv:abs/1805.10392 (2018)
13. Jadhav, A., Rajan, V.: Extractive summarization with SWAP-NET: sentences and words from alternating pointer networks. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 142–151 (2018)
14. Liu, Yang and Mirella Lapata.: Text Summarization with Pretrained Encoders. arXiv:abs/1908.08345 (2019)
15. Givchi, A., Ramezani, R., Baraani-Dastjerdi, A.: Graph-based abstractive biomedical text summarization. J. Biomed Inform. **132**, 104099 (2022). https://doi.org/10.1016/j.jbi.2022.104099.Epub 2022 Jun 11. PMID: 35700914

16. Rothe, S., Narayan, S., Severyn, A.: Leveraging pre-trained checkpoints for sequence generation tasks. Trans. Assoc. Comput. Linguist. **8**, 264–280 (2020)
17. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach. arXiv:abs/1907.11692 (2019)
18. Sharma, R., et al.: End-to-end speech summarization using restricted self-attention. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8072–8076 (2021)
19. Matsuura, K., et al.: Leveraging large text corpora for end-to-end speech summarization. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5 (2023)
20. Monteiro, R., Pernes, D.: Towards end-to-end speech-to-text summarization. arXiv:abs/2306.05432 (2023)
21. Gangi, M.A.D., et al.: MuST-C: a multilingual speech translation corpus. In: North American Chapter of the Association for Computational Linguistics (2019)
22. Lewis, Mike et al.: BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Annual Meeting of the Association for Computational Linguistics (2019)
23. Chen, D., Bolton, J., Manning, C.D.: A thorough examination of the CNN/daily mail reading comprehension task. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2358–2367 (2016)
24. Kumar, G.K., et al.: Towards building text-to-speech systems for the next billion users. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5 (2022)
25. Xue, L., et al.: mT5: a massively multilingual pre-trained text-to-text transformer. In: North American Chapter of the Association for Computational Linguistics (2020)
26. Grusky, M., Naaman, M., Artzi, Y.: Newsroom: a dataset of 1.3 million summaries with diverse extractive strategies. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 708–719 (2018)
27. Vaswani, A., et al.: Attention is all you need. In: Neural Information Processing Systems (2017)
28. Ott, M., et al.: fairseq: a fast, extensible toolkit for sequence modeling. In: North American Chapter of the Association for Computational Linguistics (2019)
29. Tang, Y., et al.: Multilingual translation from denoising pre-training. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP, 2021, pp. 3450–3466 (2021)
30. Kudugunta, S., et al.: MADLAD-400: a multilingual and document-level large audited dataset. arXiv:abs/2309.04662 (2023)
31. Seamless Communication, et al.: Seamless: multilingual expressive and streaming speech translation. arXiv:abs/2312.05187 (2023)
32. Nllb team, et al.: No Language Left Behind: Scaling Human-Centered Machine Translation. arXiv:abs/2207.04672 (2022)
33. Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**, 1401–14067 (2019)
34. Ali, A., Renals, S.: Word error rate estimation for speech recognition: e-WER. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 20–24 (2018)
35. Post, M.: A call for clarity in reporting BLEU scores. In: Conference on Machine Translation (2018)

36. Lin, C.-Y.: ROUGE: a package for automatic evaluation of summaries. In: Annual Meeting of the Association for Computational Linguistics (2004)
37. Zhang, T., et al.: BERTScore: evaluating text generation with BERT. arXiv:abs/1904.09675 (2019)
38. Zhong, M., et al.: Towards a unified multi-dimensional evaluator for text generation. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 2023–2038 (2022)

Speech and Language Resources

The ParlaSpeech Collection of Automatically Generated Speech and Text Datasets from Parliamentary Proceedings

Nikola Ljubešić[1,2(✉)], Peter Rupnik[1], and Danijel Koržinek[3]

[1] Jožef Stefan Institute, Ljubljana, Slovenia
{peter.rupnik,nikola.ljubesic}@ijs.si
[2] Faculty of Information and Communication Science, University of Ljubljana, Ljubljana, Slovenia
[3] Polish-Japanese Academy of Information Technology, Warsaw, Poland
danijel@pjwstk.edu.pl

Abstract. Recent significant improvements in speech and language technologies come both from self-supervised approaches over raw language data as well as various types of explicit supervision. To ensure high-quality processing of spoken data, the most useful type of explicit supervision is still the alignment between the speech signal and its corresponding text transcript, which is a data type that is not available for many languages. In this paper, we present our approach to building large and open speech-and-text-aligned datasets of less-resourced languages based on transcripts of parliamentary proceedings and their recordings. Our starting point are the ParlaMint comparable corpora of transcripts of parliamentary proceedings of 26 national European parliaments. In the pilot run on expanding the ParlaMint corpora with aligned publicly available recordings, we focus on three Slavic languages, namely Croatian, Polish, and Serbian. The main challenge of our approach is the lack of any global alignment between the ParlaMint texts and the available recordings, as well as the sometimes varying data order in each of the modalities, which requires a novel approach in aligning long sequences of text and audio in a large search space. The results of this pilot run are three high-quality datasets that span more than 5,000 h of speech and accompanying text transcripts. Although these datasets already make a huge difference in the availability of spoken and textual data for the three languages, we want to emphasize the potential of the presented approach in building similar datasets for many more languages.

Keywords: Spoken corpora · Parliamentary proceedings · Long speech to text alignment

1 Introduction

Although self-supervision has been shown to be the main driving force in recent drastic improvements in intelligent data processing of different data modalities,

A. Karpov and V. Delić (Eds.): SPECOM 2024, LNAI 15299, pp. 137–150, 2024.
https://doi.org/10.1007/978-3-031-77961-9_10

including image [21], video [22], text [1], speech [4], as well as different modalities [3] explicit supervision via text and speech correspondence has proven to still be the most valuable signal in developing technologies that allow processing of spoken data [20]. In this paper, we are tackling one of the more promising approaches to obtain text and speech aligned data for a large number of languages, namely through parliamentary recordings and their available manual transcripts.

1.1 Motivation

The availability of speech and text datasets differs drastically between languages, the Common Voice project [2] being a good approximation of the overall language distribution among such data: a few languages having very good coverage, some languages having decent coverage, and a long tail of languages with very limited or no coverage. The three pilot languages that we are dealing with in this paper depict the problem of the long tail very clearly. Polish, an official EU language with more than 40 million speakers, has 180 h of material in the latest version of the dataset, Serbian has 12 h, while Croatian, another official EU language with 4 million speakers, is still not present in the dataset. Croatian is not only not present in this data set, but before our efforts, there was no single open speech and text data set available for that language [15].

A convenient source of speech and the corresponding text data for official languages are parliamentary proceedings. This is because of regulations that often require the transcripts of parliamentary proceedings to be publicly available, as well as because of the frequent availability of the recordings of the proceedings as part of the public domain. The availability of speech data in the public domain is especially useful as it resolves quite a number of questions related to the biometric properties of the speech signal and the underlying privacy issues.

1.2 Prerequisites

In recent years, two iterations of the ParlaMint project [8] were funded by the CLARIN ERIC infrastructure on language resources and technologies. The main goal was to uniformly encode transcripts of parliamentary proceedings of various European parliaments. With these efforts, the availability of parliamentary transcripts for almost all official European languages has improved drastically. The current number of national parliaments covered is 26.

As part of the third iteration of the ParlaMint project, a pilot, coined ParlaSpeech, has been run with the goal of exploiting the improved availability of the textual transcripts by aligning these transcripts to the recordings of parliamentary sessions, ensuring the availability of text and speech datasets in languages not previously adequately covered with such data. For this pilot, the three already mentioned languages were chosen: Croatian, Polish and Serbian. The reasons for including exactly these three languages in the pilot were: (1) there is little to no data available for these languages, (2) there is a significant amount of transcript data available inside ParlaMint, (3) the main proponents of the

ParlaSpeech pilot have good knowledge of both the ParlaMint datasets for these languages, as well as knowledge of the languages themselves, (4) the recordings from these parliaments are available through YouTube.

Before scaling to all three languages, an initial run of ParlaSpeech was performed only for the Croatian language [15], resulting in the first publicly available text and speech dataset for the Croatian language of 1,816 h in size, together with the first ASR systems trained on a subset of available data. In this paper, we describe the second iteration of this effort, where we took the lessons learned from the initial iteration and scaled it up to three languages, with the obvious goal of scaling the approach further to even more languages in follow-up activities.

1.3 Main Challenges

While dealing with the problem of aligning parliamentary transcripts to recordings of parliamentary sessions, the following main challenges were identified: (1) parts of audio recordings are not transcribed, (2) some transcribed recordings are not released to the public, (3) parts of the recordings are transcribed with significant deviations from what has actually been said, (4) the metadata released with the recordings and the transcripts, such as the date of the session being recorded or transcribed, do not correspond, and (5) the order of texts in the transcripts does not follow the order of the events in the recording.

1.4 Similar Projects

Exploiting parliamentary data to build spoken corpora or text and speech datasets is by far not a new idea. There have been successful efforts in building such open datasets for Swiss French and German [10], Icelandic [9], Danish [12], Czech [13,14], Swiss German [17] Norwegian [24], and Finnish [28]. However, this is the first project where an approach that can be scaled to many languages is developed. A crucial component of this approach is, of course, the availability of comparable text transcripts from the ParlaMint project.

1.5 Paper Overview

The remainder of the paper is structured as follows. In Sect. 2 the problem of matching long sequences of text and speech is described, especially in light of parliamentary data and its challenges. Section 3 describes the proposed alignment procedure. Section 4 discusses post-processing decisions motivated by the release of each dataset in three flavors, described in Sect. 5, namely as (1) a FAIR repository entry, (2) a HuggingFace dataset for simplifying usage for automatic speech recognition and related tasks, and (3) a corpus in a concordancer enabling advanced search through the dataset. The paper ends with a conclusion and a description of future directions.

2 Long Speech to Text Sequence Alignment

The core of the problem discussed in the paper is aligning a long audio recording of speech to a long body of text to acquire word-level timestamps. The audio is derived from a video archive available online, and the text is a corpus derived from a human transcript of the audio made at an unknown point, usually by a government entity in accordance with local laws and traditions. The length of the video is commonly several hours, and the text usually spans a whole day's session of parliamentary proceedings. A single session is sometimes divided into several video recordings. There is no guarantee of completeness or ordering in the two sequences: there is a possibility that some information is lacking in either sequence, and the order of chunks of tokens within the text sequence could differ from the temporal order of the audio. Furthermore, the accuracy of the transcript is not ideal, because the purpose of the stenography is to create a transcript that is easy to read, rather than something that precisely depicts the audio, with all the details specific for spoken communication, such as overlapping speech and disfluencies.

2.1 Existing Approches

The idea of aligning long sequences is not new. Typical forced alignment suffers from exponential growth in complexity with respect to the length of the sequence being aligned, but there are methods to overcome these limitations. In [11], the approach was to first perform text-to-text alignment of the actual reference to the transcript perceived by the acoustic model. The acoustic model transcript is acquired using an ASR system fine-tuned to the real reference, but allowing for discrepancies through the use of an N-gram language model. The text-to-text alignment finds regions of exact matches (i.e., speech landmarks) with gaps that do not match. The matches are assumed to be correct and the mismatches are recursively re-aligned, using the same procedure, until convergence. A similar approach was used in [16] to create an alignment between audiobooks of the LibriVox project and their original text present in Project Gutenberg. That procedure was a bit more complex, as it included a more advanced ASR solution, a better language model adaptation technique, and confidence-based filtering of ASR output. The general idea remained the same.

2.2 Our Approach

This paper describes a method that is an adaptation of the above approaches with two major upgrades: (1) the utilization of a more modern end-to-end speech recognition system and (2) modifications of the text-to-text matching routine to suit the requirements of the data and the purpose of the final product. More specifically, the purpose of the older method described above is to acquire the most accurate alignment while assuming the input data to be complete and accurate. The latter method, on the other hand, simply looks for the creation of a decent corpus used to train speech recognition models - completeness of the

final product is of lesser concern. In the case of this paper, we know the data is incomplete, but we strive to achieve as much coverage as possible because the purpose is to index the data and allow further research in various settings (e.g., linguistic studies or political science studies). This is reflected in the heuristics described below. Following is a description of the complete pipeline and all its components as illustrated in Fig. 1.

3 Processing Pipeline Description

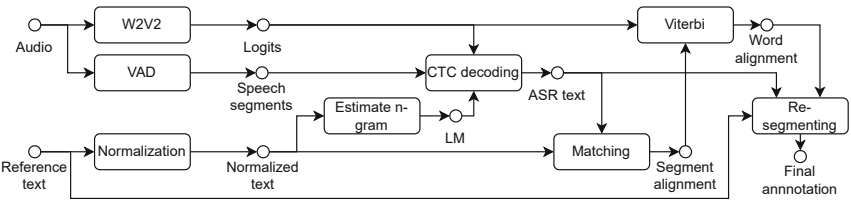


Fig. 1. Diagram of long speech to text sequence alignment pipeline for processing a single audio-text file pair. Circles are intermediary data structures. Rectangles are processes.

The procedure for aligning a large dataset begins with a collection of audio files and a long text corpus. We iterate by audio files and need to find a chunk of text that is reasonably long to cover the contents of that audio recording. The corpus is divided into sections with corresponding metadata, but in our experience, there is frequently a mismatch between the recording and transcript metadata. That is why we utilize a statistical analysis - ratio of n-gram coverage, comparing the reference transcripts to ASR outputs of each file to see what is the best match. Sometimes, several transcripts are matched to a single recording, and vice versa. For the remainder of this section, we will assume that the matching between the recordings and transcript is present and start the pipeline description from a single audio recording and transcript covering the contents of that file.

3.1 Audio Processing

The pipeline for processing a single recording and transcription pair is designed to facilitate mass data processing, so the intermediary results are saved in a cache for reuse in multiple stages along the way. Such is the case with speech processing – instead of performing ASR in one go, we first compute the Wav2Vec2-XLS-R (W2V2) model [7] logarithmic likelihoods (logits) into a file and then use it both for ASR decoding and Viterbi alignment, which occurs later in the process.

Except for the calculation of W2V2 logits, we also use Voice Activity Detection (VAD) to figure out which parts of the recordings contain speech that is

likely being transcribed by the human transcribers. We use the pyannote [6] package to extract speech segments and then remove segments with an energy level below −45 dB RMS. This is to remove most of the background conversations that are generally not transcribed. Both W2V2 logits extraction and VAD processing are computed on the GPU and caching their outputs allows for optimal use of the hardware. All other processes are computed on the CPU.

3.2 Text Pre-processing

At the same time, we normalize the reference text. Depending on the W2V2 model, the output of ASR may or may not contain digits, but usually does not contain most of the symbols, punctuation or capitalization. To better accommodate the matching of the ASR output to the human transcript, we need to normalize the human transcript to remove any punctuation and capitalization, as well as convert any symbolic text into its pronounced form. This is a somewhat language-dependent procedure, and although it is well researched [5], we had to rely on custom rule-based solutions for the languages and corpora being prepared in this paper. Even though it is a common procedure in, e.g., text-to-speech software, no high-quality open-source solutions existed for the languages analyzed at the time of performing the research.

3.3 Language Modeling and Speech Recognition

The normalized text is also used to prepare the language model (LM) for the ASR decoding phase of the procedure. We tried to train the model only on the text in the transcript being processed, but we obtained much better results by combining all the transcripts and creating a single LM for all the files being processed. This is most likely due to the better statistics acquired with a larger quantity of text. We used the SRILM toolkit [25] to train a Knesser-Ney discounted 3-gram model with interpolation. We then use the pyctcdecode[1] package to generate the ASR output text based on the W2V2 logits and the language model. The VAD output is also used at this stage to determine which parts of the audio should be processed.

3.4 Matching of Automatic to Reference Text

The next step in the pipeline is to match the generated ASR output with the normalized reference text. The procedure, as illustrated in Fig. 2, starts by looking for potential matches between the two text sequences using a word histogram of a sliding window. Next, each match is evaluated using the Levenshtein distance to find the best match. The nonmatching prefix and suffix (so any insertion or deletion at the start and end of sequence) is rejected, and everything else is treated as the final match. Depending on the chosen thresholds, this method leaves gaps in the final result. We then try to force the matches in those gaps

[1] https://github.com/kensho-technologies/pyctcdecode.

by using the Levenshtein comparison again. Sometimes this still leaves gaps, particularly if they are large or if the order of the sequences does not match. In that case, we try and repeat the whole procedure outlined above, starting with the histogram-based search of the whole reference, but only within the remaining ASR gaps. The statistics of coverage (i.e. how many words are matched in either sequence) are computed along each phase of the procedure, which helps in tuning the thresholds and other hyperparameters.

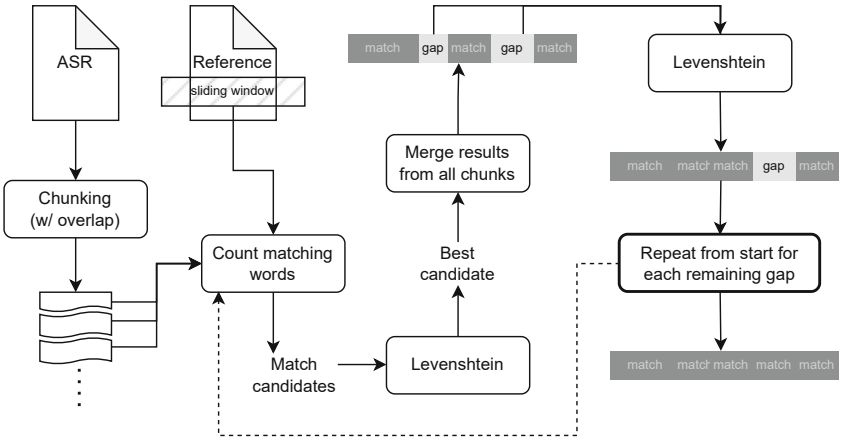


Fig. 2. Illustration of the matching algorithm. The purpose is to find portions of the reference that match the ASR output. Sequence and accuracy is not guaranteed.

The output of the matching is a list of audio segments and their matching reference text. These segments can span many words, so in order to obtain time offsets for each individual word, we need to re-align the audio to the actual reference text rather than the ASR output obtained earlier. For this, we use a simple Viterbi forced alignment algorithm [19] to match the character sequence to the output of the W2V2 model. One feature of the W2V2 model is the presence of word delimiter tokens in the output: to obtain word level offsets, it is sufficient to look for the location of the word delimiters within the aligned sequence.

3.5 Post-processing

In the final step of the pipeline, all the information from previous stages is combined to create one coherent, aligned annotation of the file. We use a mapping between the original and normalized reference text to project the time offsets onto the original un-normalized token sequence. We also add the ASR output, temporally aligned with the reference sequence above. This is both for visualizing possible errors in human transcripts (by calculating the word error rate between the ASR output and the reference text) and for providing automatic transcription for parts of audio that human transcribers did not transcribe. Finally, we

map everything into chunks that were used in the original reference text corpus. Each chunk of the original corpus contains a unique ID, which allows us to combine our acoustic time annotation with other forms of annotation and metadata present in the original text corpus.

4 Segmentation and Filtering

In this section, we describe the further segmentation and accompanying filtering from our text and audio alignment process, presented in the previous section. This segmentation and filtering are necessary to generate the datasets and corpora that we are currently considering to be most useful for downstream usage. We describe such three downstream cases in the next section.

The output of the alignment process from the previous section are JSON files, one per each original audio file, consisting of three types of entries: (1) an ASR transcription of the part of the audio file that could not be matched to any part of the ParlaMint corpus, (2) a speech transcript from the ParlaMint corpus that could not have been aligned with this part of this audio file, but its surrounding transcripts could, and (3) the speech transcript from the ParlaMint corpus together with the ASR transcription that was matched to that particular ParlaMint transcript, along with a list of predicted word alignments consisting of character offsets (referencing to the ParlaMint speech transcript) and millisecond offsets (referencing to the audio recording).

To obtain streamlined datasets from this matching output, which would be better suited for downstream applications such as aligned text and speech datasets for training automatic speech recognition, or spoken corpora for linguistic purposes, a series of additional filterings and segmentations have been performed.

The first filtering iteration was performed at the level of speeches, removing all ParlaMint speech transcripts without audio alignment, as well as speeches with alignment, but with an estimated character error rate between the ParlaMint transcript and the aligned ASR transcription higher than or equal to 60%. The aim of this filter was to remove nonaligned portions of the data, as well as those portions where even partial alignments are very questionable.

In the next step, each speech transcript was segmented into sentences to filter out parts of the speeches with a lack of correspondence. For each sentence covered by word alignment information, the character error rate was calculated again between the ParlaMint transcript and the ASR transcription, and all sentences with a character error rate greater than 10% were filtered out. With this filter, sentences with deviation between the spoken signal and the transcript have been discarded.

A final filtering at the sentence level was performed in cases where the ratio of the length of the audio in milliseconds and the length of the transcript in characters were greater than 0.2. Namely, in some cases, the alignment process matched a sentence to part of the audio with longer or shorter breaks in the work of the parliament, making the audio unrealistically longer than what would

be expected given the length of the transcript. Given that these breaks are mostly muted audio, such imperfections could not have been filtered out with the previous filters based on the character error rates.

All three filtering thresholds were defined by inspecting samples of data and manually identifying reasonable cut-off points.

To estimate the yield rate of the whole matching and filtering procedure, we calculated the percentage of ASR transcriptions that were successfully matched to the ParlaMint transcripts after all filterings. We performed this yield estimation on the Croatian data as we are rather positive that the recordings are to the most part covered with the ParlaMint transcripts. The yield rate for matching the available audio information with the textual transcript in this particular case was 74%. Given our experience with the data, which also includes manual analyses of the non-aligned ASR transcriptions, the main reasons, in order of prevalence, for parts of the spoken content not being aligned to the ParlaMint transcripts are: (1) speech not being transcribed within the parliament, (2) transcripts differing from the spoken word, (3) ASR errors, and (4) matching errors.

5 The Dataset Releases

In this section, we describe the three encodings of our datasets aimed at the specific downstream use cases: master CLARIN.SI FAIR (findable, accessible, interoperable, reusable) repository entries aimed at archiving all available information, HuggingFace datasets practical for using our data on tasks such as training automatic speech recognition and various speech classification models, and linguistically annotated corpora that allow complex linguistically informed searches through the datasets.

5.1 FAIR Repository Entries

As the master release of the produced datasets we have prepared jsonl files, each line covering a single sentence from the ParlaMint corpus, with a reference to the corresponding flac audio file, and word-level alignment containing character offset and millisecond offset information.

In addition to the spoken and textual content and their alignment information, a significant amount of metadata present in the ParlaMint corpus has also been included in these datasets. Inter alia, the following information on the speaker was included: the role of the speaker (are they speaking as a chairperson or an MP), the party they belong to, the political orientation of the party, whether the party is in coalition or opposition at the time of the speech, and the gender and their year of birth of the speaker.

We publish such prepared datasets on the FAIR (findable, accessible, interoperable, reusable) repository of CLARIN.SI,[2], the Slovenian national node of the CLARIN ERIC infrastructure on language resources and technologies.[3]

[2] https://www.clarin.si/repository/xmlui/.
[3] https://www.clarin.eu.

The statistics on the size of each of these releases are presented in Table 1. From the presented numbers, it is obvious that the Croatian dataset is by far the largest. While for both Croatian and Polish, all available data were processed, for the Serbian dataset, only 1000 audio files out of almost 4500 files have been processed by now. We hope that we will be able to further expand the Serbian dataset in one of the next ParlaSpeech data collection releases.

Table 1. The statistics on the size of the three ParlaSpeech datasets.

corpus	HR	PL	RS
size (GB)	179.0	60.8	57.7
duration (h)	3110.39	1009.82	896.22
sentences	922 679	535 465	290 778
words	24 755 742	7 515 333	7 024 293
characters	150 970 948	52 724 103	42 638 259
median sentence (s)	9.62	4.94	8.74

The repository entries can be accessed through persistent identifiers for Croatian[4], Polish[5] and Serbian[6].

5.2 HuggingFace Datasets

The second release of the dataset is through the HuggingFace Datasets Hub,[7], which allows technical users to gain access to all three data sets using just a few lines of code.

The data sets are again available separately for Croatian[8], Polish[9], and Serbian[10].

Given that parts of the available speaker metadata were included in the HuggingFace datasets, such as the gender, age, party affiliation, whether the speaker was in coalition or opposition during the speech, the data are useful for so much more than just automatic speech recognition. We are looking forward to all the interesting use cases that this data availability will produce.

5.3 Spoken Corpora via Concordancer

Finally, the third availability of the dataset is aimed at the use of linguists and phoneticians. Each dataset has been made available through the CLARIN.SI

[4] http://hdl.handle.net/11356/1914.
[5] http://hdl.handle.net/11356/1686.
[6] http://hdl.handle.net/11356/1834.
[7] https://huggingface.co/datasets.
[8] https://huggingface.co/datasets/classla/ParlaSpeech-HR.
[9] https://huggingface.co/datasets/classla/ParlaSpeech-PL.
[10] https://huggingface.co/datasets/classla/ParlaSpeech-RS.

concordancer[11] with the help of which the textual transcript can be searched with the Corpus Query Language, and the recording of each search result can be played back.

To enable more detailed searches using the Corpus Query Language, each of the sentences in the corpus was linguistically annotated, splitting each sentence into words and annotating each word with the part of speech, the morphosyntactic features, and the lemma of the word. For Croatian and Serbian, the CLASSLA-Stanza tool [27] was used, while for Polish we applied the Stanza tool [18].

To make the playback of the recordings as user-friendly as possible, with the median length of sentence recordings between 5 and 10 s, depending on the language, for this release we have performed another segmentation of the data with the aim of obtaining recordings in length between 3 and 6 s. If a researcher requires recording of the entire sentence, it can still be accessed in the metadata of each sentence.

The availability of the datasets through concordancers is again separated by language, having a separate corpus for Croatian[12], Polish[13], and Serbian[14].

The intended use of the concordancer is to simplify linguists and phoneticians' identification of linguistic patterns that they are interested in, and accessing their recordings that can be further processed in specific tools such as Praat [26], or Exmaralda [23].

An example of the result of a search for the noun "tehnologija" with a preceding adjective in the Croatian corpus is presented in Fig. 3. The sought phrase is colored red, the surrounding context is colored black, with an icon for playing the recording of the phrase to the right, and a link to the speaker and sentence metadata to the left.

6 Conclusion

In this paper, we have presented a robust and scalable approach to aligning thousands of hours of recordings of parliamentary proceedings with their manual transcripts released by the parliaments. This process includes a series of challenges, the biggest ones being the non-correspondence in either the data coverage or the data order in any of the two data modalities. Furthermore, deviations in the transcripts from the spoken words are a rather frequent phenomenon. The reason for these deviations is that the transcripts are not aimed at performing linguistic research or speech technology development, but to ensure availability of the parliamentary discussions for the general public.

We have described three complementary approaches to releasing the final datasets - (1) a complete, master release through a FAIR repository to ensure maximum availability and reusability of the data, (2) the opportunistic release

[11] https://www.clarin.si/ske/.
[12] https://www.clarin.si/ske/#dashboard?corpname=parlaspeech_hr.
[13] https://www.clarin.si/ske/#dashboard?corpname=parlaspeech_pl.
[14] https://www.clarin.si/ske/#dashboard?corpname=parlaspeech_rs.

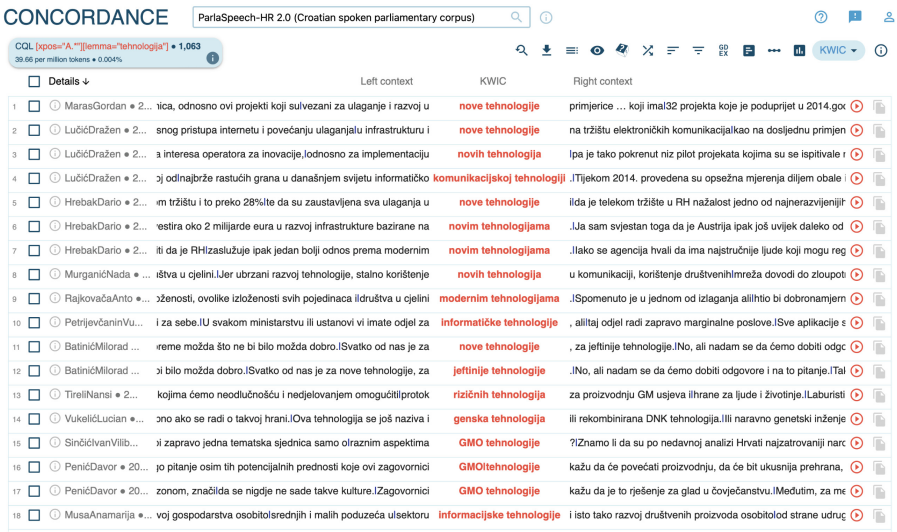


Fig. 3. Example of a search result on the noun "tehnologija" with a preceding adjective in the concordancer of the Croatian corpus. The recording can be accessed to the right, the metadata to the left.

through the HuggingFace Datasets Hub to simplify speech technology development, not only on the problem of automatic speech recognition, but also additional tasks such as demographic prediction due to availability of rich metadata, and (3) the release in form of spoken corpora available through a linguistic concordancer, which allows linguists to search the transcripts, enriched with linguistic features such as part of speech, lemma, and morphosyntactic features, as well as to listen or to retrieve the recordings of their search results.

There are limitations to our work. The first is on the side of the data that come from the rather limited parliamentary domain and, even more, they are filtered by the correspondence between the recording and the transcript, which removes all the content where transcribers were performing stronger edits, such as disagreements, disfluencies, etc. The second is on the side of the method that requires an at least partially functioning speech encoder and related ASR system, but also the availability of the transcripts and the recordings of the parliamentary sessions.

With the presented results of more than 5,000 h of corresponding speech and text in three less-resourced Slavic languages, we are of the opinion that we have just scratched the surface of what the ParlaSpeech concept can bring to the research community. We will primarily focus on adding additional languages to the ParlaSpeech collection, as there is a significant number of the current ParlaMint languages that would immensely profit from the availability of ParlaSpeech data for that language. In parallel with that, we will use the ParlaSpeech data in both technology development and linguistic and

communication research, especially looking out for various types of biases in the data themselves, as well as biases that we have introduced with our matching and filtering procedures.

Acknowledgments. The research presented in this paper was conducted within the research project "ParlaMint: Towards Comparable Parliamentary Corpora" funded by CLARIN ERIC. The research was also co-funded by the research project titled "Basic Research for the Development of Spoken Language Resources and Speech Technologies for the Slovenian Language" (J7-4642), and withing the research programme "Language resources and technologies for Slovene" (P6-0411), both funded by the Slovenian Research and Innovation Agency (ARIS).

References

1. Achiam, J., et al.: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Ardila, R., et al.: Common voice: a massively-multilingual speech corpus. arXiv preprint arXiv:1912.06670 (2019)
3. Baevski, A., Hsu, W.N., Xu, Q., Babu, A., Gu, J., Auli, M.: Data2vec: a general framework for self-supervised learning in speech, vision and language. In: International Conference on Machine Learning, pp. 1298–1312. PMLR (2022)
4. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: a framework for self-supervised learning of speech representations. In: Advances in Neural Information Processing Systems, vol. 33, 12449–12460 (2020)
5. Bakhturina, E., Zhang, Y., Ginsburg, B.: Shallow fusion of weighted finite-state transducer and language model for text normalization (2022). https://arxiv.org/abs/2203.15917
6. Bredin, H.: pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In: Proceedings of the INTERSPEECH 2023 (2023)
7. Conneau, A., Baevski, A., Collobert, R., Mohamed, A., Auli, M.: Unsupervised cross-lingual representation learning for speech recognition (2020). https://arxiv.org/abs/2006.13979
8. Erjavec, T., et al.: The parlamint corpora of parliamentary proceedings. Lang. Resour. Eval. **57**(1), 415–448 (2023)
9. Helgadóttir, I.R., Kjaran, R., Nikulásdóttir, A.B., Guðnason, J.: Building an ASR corpus using althingi's parliamentary speeches. In: Interspeech, pp. 2163–2167 (2017)
10. Imseng, D., Bourlard, H., Caesar, H., Garner, P.N., Lecorvé, G., Nanchen, A.: MediaParl: bilingual mixed language accented speech database. In: 2012 IEEE spoken language technology workshop (SLT), pp. 263–268. IEEE (2012)
11. Katsamanis, A., Black, M., Georgiou, P.G., Goldstein, L., Narayanan, S.: SailAlign: robust long speech-text alignment. In: Proceedings of Workshop on New Tools and Methods for Very-Large Scale Phonetics Research, vol. 1 (2011)
12. Kirkedal, A., Stepanović, M., Plank, B.: Ft Speech: danish parliament speech corpus. arXiv preprint arXiv:2005.12368 (2020)
13. Kopp, M., Stankov, V., Krůza, J.O., Straňák, P., Bojar, O.: ParCzech 3.0: a large czech speech corpus with rich metadata. In: Ekštein, K., Pártl, F., Konopík, M. (eds.) TSD 2021. LNCS (LNAI), vol. 12848, pp. 293–304. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-83527-9_25

14. Kratochvíl, J., Polák, P., Bojar, O.: Large corpus of czech parliament plenary hearings. In: Proceedings of the Twelfth Language Resources and Evaluation Conference, pp. 6363–6367 (2020)
15. Ljubešić, N., Koržinek, D., Rupnik, P., Jazbec, I.P.: ParlaSpeech-HR-a freely available ASR dataset for croatian bootstrapped from the ParlaMint corpus. In: Proceedings of the workshop ParlaCLARIN III within the 13th language resources and evaluation Conference, pp. 111–116 (2022)
16. Panayotov, V., Chen, G., Povey, D., Khudanpur, S.: Librispeech: An ASR corpus based on public domain audio books. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5206–5210 (2015). https://doi.org/10.1109/ICASSP.2015.7178964, https://ieeexplore.ieee.org/document/7178964
17. Plüss, M., et al.: SDS-200: A swiss German speech to standard German text corpus. arXiv preprint arXiv:2205.09501 (2022)
18. Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C.D.: Stanza: A Python natural language processing toolkit for many human languages. arXiv preprint arXiv:2003.07082 (2020)
19. Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE **77**, 257–286 (1989). https://api.semanticscholar.org/CorpusID:13618539
20. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: International Conference on Machine Learning, pp. 28492–28518. PMLR (2023)
21. Rani, V., Nabi, S.T., Kumar, M., Mittal, A., Kumar, K.: Self-supervised learning: a succinct review. Arch. Comput. Methods Eng. **30**(4), 2761–2775 (2023)
22. Schiappa, M.C., Rawat, Y.S., Shah, M.: Self-supervised learning for videos: a survey. ACM Comput. Surv. **55**(13s), 1–37 (2023)
23. Schmidt, T., Wörner, K.: EXMARaLDA (2014)
24. Solberg, P.E., Ortiz, P.: The norwegian parliamentary speech corpus. arXiv preprint arXiv:2201.10881 (2022)
25. Stolcke, A., et al.: SRILM-an extensible language modeling toolkit. In: Interspeech, vol. 2002, p. 2002 (2002)
26. Styler, W.: Using praat for linguistic research. University of Colorado at Boulder Phonetics Lab (2013)
27. Terčon, L., Ljubešić, N.: Classla-stanza: the next step for linguistic processing of south slavic languages. arXiv preprint arXiv:2308.04255 (2023)
28. Virkkunen, A., Rouhe, A., Phan, N., Kurimo, M.: Finnish parliament ASR corpus: analysis, benchmarks and statistics. Lang. Resour. Eval. **57**(4), 1645–1670 (2023)

ESC Corpus of Spoken Russian: Everyday Student Conversations Captured Through Continuous Speech Recording in Natural Communicative Environments

Tatiana Y. Sherstinova(✉) and Irina Petrova

HSE University, Saint Petersburg, Russia
{tsherstinova,ia.petrova}@hse.ru

Abstract. This article describes the methodology for creating a new resource of everyday Russian speech, based on audio recordings made by student volunteers over the course of an entire day in natural communication settings (at home, in the university, at the café, in the fitness club, etc.). The precursor to this corpus is the well-known ORD corpus, or the "One Day of Speech" corpus, for which recordings were made from 2007 to 2016. Since the ORD recordings were made, certain changes have occurred in Russian spoken language, particularly noticeable in the speech of young people at the lexical level. The creation of the new speech resource aims to capture this linguistic snapshot to identify new colloquial vocabulary, as well as new meanings and connotations of known language units. The new recordings of everyday spoken language will supplement the empirical material of the ORD corpus and provide a foundation for various scientific, theoretical, and practical endeavors. The article details the methodology for creating the Everyday Student Conversations (ESC) corpus, highlights its differences from the ORD corpus, and provides current ESC corpus statistics.

Keywords: Corpus Linguistics · Everyday Spoken Russian · Oral Discourse · Speech Corpus · Student Speech · Field Linguistics · Day-long Recording

1 Introduction

This article describes the methodology of creating a new multimedia language resource—the ESC Corpus, designed for the study of everyday Russian spoken communication and spontaneous speech used in both casual and professional settings. The abbreviation ESC stands for "Everyday Student Conversations"[1], indicating that the core of this resource is the communication among young people, primarily students. The development of this resource was initiated by the Laboratory for Language Convergence at the National Research University Higher School of Economics in Saint Petersburg in 2023.

[1] In Russian publications, the corpus is also referred to as *KURS*, or the "Corpus of Students' and Youth's Spoken Language".

A. Karpov and V. Delić (Eds.): SPECOM 2024, LNAI 15299, pp. 151–162, 2025.
https://doi.org/10.1007/978-3-031-77961-9_11

The ESC Corpus is a resource based on recordings of everyday speech interactions made in field conditions. The immediate predecessor of the ESC corpus is the Russian ORD corpus of everyday conversations, also known as "One Day of Speech" corpus [1; 2], which was created using a continuous audio recording methodology specifically developed by its authors. Volunteers who participated in the corpus recording received a voice recorder, which they carried with them throughout their regular day, from the moment they woke up until they went to bed. The voice recorder accompanied the participants throughout the day, capturing all their speech interactions in various communicative settings—at home, at the university, at work, in cafés, in stores, in service centers, etc. [6]. A similar recording methodology was previously used in collecting the demographic sample of the British National Corpus for English [3] and the Japanese ESP Corpus, which is part of the large JST/CREST project dedicated to processing emotional speech [4]. This approach allowed for the collection of speech samples from the real language environment—essentially, *field* recordings. The demographic balancing of informants enabled the collection of speech material from representatives of different social groups, thereby broadening the spectrum of communication conditions. The materials of the ORD corpus served as the basis for a series of scientific studies ranging from describing the vocabulary, grammar, and pragmatics of everyday speech to investigating its sociolinguistic variability and examining communicative scenarios and other aspects of natural speech communication.

Since the ORD recordings were made in 2007–2016 [5], certain changes have occurred in Russian spoken language, particularly noticeable in the speech of young people at the lexical level. The creation of the new ESC speech resource aims to capture this linguistic snapshot to identify new colloquial vocabulary, as well as new meanings and connotations of known language units.

The ORD corpus became the prototype for the development of the new resource. Processing audio recordings collected in its manner poses a non-trivial challenge due to several factors inherent in field recordings of natural speech, as compared to "laboratory" recordings made in a professional studio (or even over the phone, as is often done for speech corpora). These factors include:

1 the presence of a significant amount of background noise (which can sometimes overshadow the speech signal depending on the specific situation);
2 significant variation in the signal level when the distance between the speaker and the recorder changes (e.g., when the recording device is placed on a table);
3 the informal nature of everyday conversations, which are generally not intended to be listened to by anyone other than the participants themselves (certain conversation fragments may be completely incomprehensible to linguists attempting to transcribe and analyze the recorded speech);
4 the virtually unlimited thematic range of conversations, which can include both universally understood "weather talk" and very narrow topics understandable only to a select group of professionals;
5 the number of conversation participants can change during the course of the conversation and can also be practically unlimited;

6 speech overlap (i.e., simultaneous speech production by several interlocutors) in natural communicative situations significantly complicates the processing of everyday field conversations.

Additionally, as with any audio resource, obtaining transcripts—written versions of spoken speech—poses a separate challenge, allowing the corpus to be used as a collection of written texts. These factors explain why there is still a shortage of representative audio resources dedicated to everyday language.

The creation of a new corpus aims to expand the empirical base for both theoretical linguistic research on spoken language and for solving practical tasks related to the development of speech systems and artificial intelligence (AI) agents that use natural conversational speech familiar to humans. The need for a new spoken language resource arises from the constant evolution of everyday speech, primarily at the lexical level, especially in the speech of the younger generation. Currently, researchers and developers have access only to the ORD corpus recordings collected from 2007 to 2016. Since then, certain changes have occurred in Russian spoken language, particularly noticeable in the speech of young people at the lexical level. Moreover, since the last ORD recordings, technical capabilities in corpus linguistics have improved, including automatic transcription and audio processing, which can be utilized in creating the new corpus.

The new ESC corpus of everyday Russian speech is intended to address the following tasks:

- To study everyday communication language, vocabulary, grammar, and pragmatics of spoken speech (both formal and informal);
- To create a "sounding memory" of our time and collect audio recordings that reflect the "sound portrait" of the era;
- To model speech communication in various communicative situations, necessary for studying the social, psychological, and cultural aspects of everyday language, as well as for solving practical tasks in teaching communication skills to people and robotic systems;
- To configure and test speech synthesis and recognition systems and other applications, closely approximating real-life conditions, associated with the development of AI systems that understand language in its natural form and use it for sound communication.

The article details the methodology for creating the Everyday Student Conversations (ESC) corpus, highlights its differences from the ORD corpus, and provides current ESC corpus statistics.

2 Methodology for Compiling the ESC Corpus

The ESC corpus is essentially a conceptual and methodological continuation of the ORD corpus. The work on creating the ESC corpus is based on the ongoing audio recording methodology developed by the authors of the ORD corpus [1]; [5]. Volunteers wishing to participate in the corpus recording receive a voice recorder, which they carry with them throughout their regular day. This approach allows for the collection of speech samples from the real language environment.

2.1 Methodology of Audio Recording Collection

Who can Become an Informant-Volunteer for the ESC Corpus Recording? As indicated by the corpus name, "Everyday Student Conversations", the core of the resource is student speech. This name was given to the corpus because it is being developed in a university environment, and the majority of its respondents are students. However, as noted in the introduction, we are also interested in the speech of young people in general, as it is in the speech of this age group that new linguistic phenomena and neologisms can be expected [7]. Therefore, contributions to the corpus from young people who are not students are also welcomed.

In different countries and international organizations, the category *youth* encompasses various age groups. For the purposes of this corpus, we define the youth age group as individuals aged 18 to 24, which is typically the age range for college and university students in Russia. Additionally, we record the speech of young adults aged 25 to 35.

Moreover, we have decided to simultaneously collect speech from volunteers of older age groups (36 and above), thus forming the basis for a sub-corpus titled "Everyday Adult Conversations". We do not impose restrictions on occupation, primary place of residence, or other social characteristics of the recording participants. Currently, anyone over the age of 18, including those for whom Russian is not a native language, can participate in the ESC corpus recordings.

It is evident that this approach does not immediately contribute to a balanced corpus material. However, it allows us to preserve unique recordings from a wide range of speakers and, in the future, scale the resource to a national corpus of everyday spoken Russian.

What is Recorded? The recordings must contain communication in Russian[2]. The primary unit of recording obtained from informants remains their "speech day", that is, a collection of audio files reflecting their speech communication from the moment they wake up in the morning until they go to bed at night. As with the ORD recordings, we ask volunteers to choose the most suitable days for recording themselves. If, during the day, situations arise whose recordings the experiment participants are unwilling to submit to the corpus for any reason, these fragments are deleted[3]. As with the ORD corpus recordings, the essential condition is the informed consent of all primary participants in the conversation to the recording being made. We ask informants to notify their potential interlocutors about the upcoming recording the day before, so that the conversation topics do not predominantly revolve around the details of data collection for the corpus.

Unlike the ORD, in collecting data for the new corpus, we decided to allow recording not only the entire speech day but also individual speech situations (e.g., "birthday celebration" or "conversation with a friend") as well as multiple entire days. This approach increases the overall amount of speech material and the diversity of informants represented in the corpus, enhancing its statistical representativeness.

[2] The presence of speech fragments in any other languages is allowed in the recordings, but currently, only Russian speech is being processed.

[3] Experience shows that ORD volunteers rarely use this option: in cases where they anticipate situations, they deem undesirable to record, respondents prefer to pause the voice recorder.

In addition to face-to-face communication, the corpus also includes the informants' conversations over the phone, via other mobile devices, and online applications.

How is the Recording Conducted? The recording methodology remains almost unchanged compared to the material collection for the ORD corpus. The recording is done using a professional digital recorder, which can be either worn around the informant's neck while they move around or placed stationary on a table during indoor recordings. The set of recording devices has been updated[4]. The format of the original recordings is WAV, stereo, 16-bit, 44,100 Hz.

Additionally, considering the significant improvement in the quality of recordings made with personal mobile devices, it was decided to allow informants to use their own smartphones for recording. The rules for preferring continuous recordings remain unchanged. Test experiments have shown that high-quality speech material can also be obtained when recording with a modern phone.

Each participant who agrees to be recorded receives a kit that includes the recording device, additional accessories (rechargeable batteries, charger, external microphone), and an informant's memo. The memo includes instructions on how to use the recorder and the main recording rules, such as: "The recording is conducted for 12–16 h (from morning to evening) with minimal interference from the participant in the recording process. The recorder is only stopped to change power sources (approximately every 9 h)". When the informant speaks on the phone or through voice messengers, the instructions suggest using the speakerphone whenever possible so that the interlocutor's speech is also recorded.

To allow sociolinguistic research on the corpus material, we ask each informant to fill out a sociological questionnaire, modeled after the one proposed by the authors of the ORD [5]. This questionnaire includes the following characteristics: 1) gender; 2) age; 3) place of birth; 4) place of residence; 5) experience of living in a region different from the current place of residence; 6) native language; 7) other languages spoken by the informant; 8) parents' profession; 9) education level; 10) acquired specialty; 11) work experience, if any. The new item "parents' profession" was introduced instead of the question "social background", which often caused confusion among ORD informants. Similar information is also collected for all primary communicants.

Furthermore, the informant's questionnaire includes consent to transfer data to the ESC corpus for scientific purposes and consent to process personal data. Additionally, the questionnaire contains an item about the informant's agreement/disagreement on the open publication of the speech material obtained from them, with the following possible responses: "Yes, the recording can be published in full", "I do not mind publishing most of the recording", "No, the audio recording cannot be published". The informant's and primary communicants' questionnaires are completed online using a cloud service. The use of the corpus implies the anonymity of the data presented—unique codes are used in place of informants' names. When publishing transcripts, all personal information is expected to be anonymized. Recordings that informants agree to publish openly will also be reviewed for the presence of sensitive personal information and anonymized as needed.

[4] Currently, the recording is primarily done using Roland R09-HR, Zoom H1n, and Tascam DR-05X voice recorders.

During the recording process, participants are required to keep a *Speech Day Diary*, briefly noting communicative events (with whom, during which period, and under what circumstances the informant communicates throughout the day). Currently, this diary is implemented via a Telegram bot.

Informants are also asked to complete a psychological test – the Five-Factor Personality Questionnaire 5PFQ [8]; [9]. This test measures the degree of expression of each of the five factors of the "Big Five": extraversion, agreeableness, conscientiousness, emotional stability, and openness to new experience [10].

In addition to the psychological test, informants are given a test to determine their passive vocabulary [11], based on J. Read's statistical approach [12]. This approach assumes that the probability of a respondent knowing words used in the language with equal frequency is approximately the same. This allows for testing the informant's knowledge not of all the words in the language, which is inherently impossible, but of only a small number of specially selected test words, each representing a group of words of approximately equal frequency [13]. This test is also conducted online.

Thus, it was decided to completely abandon paper documentation in creating the corpus: all communication with informants, their questionnaires, and tests are conducted online.

The corpus recording is conducted anonymously. After receiving audio files from informants, their data are encrypted. Each participant is assigned a code in the format like *AF00n*, where the first letter *A* indicates that it is informant's code, the second letter denotes gender *F* or *M*. This is followed by the serial code of the participant.

2.2 Principles of Speech Data Processing

Segmentation of the Original Recordings into Macro Episodes. The multi-hour audio files obtained from the informants are listened to by experts. They remove long pauses or noise fragments that do not contain useful information. After that, the files are segmented into macro episodes, which are large episodes united by the place of communication, its conditions, and participants (for example, "breakfast with the family", "work meeting", "working with clients in the office", "shopping in the store", etc.) [14]. Each episode is annotated using the formal description methodology for macro episodes. Typically, the duration of a macro episode ranges from 15 to 40 min. Each episode is given a code name that reflects the name of the corpus, the unique informant code, and the episode number (e.g., *escAF001–01*). The output file format, which predominantly contains speech, is mono, 16-bit, 22,050 Hz.

The annotation of macro episodes is performed using the ORD methodology [ibid.]: the main type of communication ("private/domestic conversation", "professional conversation", "public speech", etc.), communication conditions (e.g., "telephone conversation", "meal", "while shopping", etc.) are identified. Additionally, tags indicate the presence of monologic speech, singing, reading, conflict situations, and other features of the communicative situation.

Furthermore, for each macro episode, the social role of the informant is noted— relative (husband, wife, daughter, brother, etc.), client or service representative, teacher or student, colleague; it is also noted when a person is talking to themselves. In addition to

the informant's role, their interlocutors and their relationship to the informant are indicated. The place of communication is also marked: home, office, educational institution, clinic, street, etc.

Audio files corresponding to macro episodes are the primary unit of description in the corpus. Simultaneously with their annotation, the expert notes the quality of the speech signal in terms of noise: 1—high-quality recording with minimal background noise, allowing phonetic research on this material; 2—relatively good quality recording; 3—average recording quality; 4—very noisy recordings.

Preparation of Transcripts of Audio Recordings. After segmentation, the audio recordings are transcribed to obtain transcripts.

Since speech recognition tools have significantly improved recently, it was decided to use modern speech recognition models for obtaining transcripts when creating the new corpus of everyday speech. During the transcription phase, the decision was made to abandon ELAN [15], which is used for ORD transcribing in favor of using tools for automatic diarization (segmenting the speech stream into parts corresponding to individual speakers) and automatic speech recognition with subsequent manual correction. For pilot experiments, two models were used—an acoustic model developed by the company NTR [16] and the multilingual neural network Whisper [17], created by OpenAI, the developer of ChatGPT [18]. The results of the test transcriptions of "speech days" are described in the works of [19–21].

The automatically obtained transcripts are then manually corrected. Experts fix program errors related to both incorrect recognition of acoustic signals at the lexical level and incorrect attribution of speakers. The corrected transcripts are added to the ESC speech corpus and used for further fine-tuning of the recognition models used.

Once the transcripts of the audio recordings are obtained, any tools for automatic text analysis can be used—tokenization, morphological and syntactic analysis, as well as specialized annotation.

2.3 The Main Differences in Data Collection and Processing Methodology Between the ESC and ORD Corpora

Despite the overall methodology for data collection being borrowed from the ORD corpus, there are several key differences in creating the ESC corpus:

- Complete abandonment of paper documentation: all questionnaires, instructions, and consents are filled out electronically online.
- Allowing the recording of one's "speech day" on a phone that supports sufficiently good recording quality.
- The corpus accepts not only recordings of entire speech days but also their individual fragments and even specific conversations.
- Transcription of audio recordings is performed in a semi-automatic mode: initially, rough transcriptions are done using speech recognition systems, which are then reviewed and corrected by experts[5].

[5] Previously, during the creation of the ORD corpus, transcriptions were done manually by experts using the ELAN program (http://tla.mpi.nl/tools/tla-tools/elan). Transcribing one minute of

3 Current Statistical Characteristics of the ESC Corpus

The first audio recordings for the new corpus were obtained in the spring of 2023. At this research stage, only students from the "Philology" program at the HSE University in Saint Petersburg were invited to participate as the project volunteers to refine the material collection and data processing methodology. Starting in 2024, the pool of volunteer candidates was expanded beyond the university.

At the time of writing, speech material has been obtained from 70 volunteers. The total duration of the original recordings is 900 h; 308 of these were recorded on a mobile phone's built-in recorder, while the rest were recorded on professional audio recording devices. The average duration of speech data obtained from an informant is 12.9 h, with the maximum recording duration from a single person being 62 h.

Segmentation of speech days into macro episodes and removal of fragments not containing useful signals have been completed for 42 informants (550 h of recording). As a result, 933 macro episodes with a total duration of 380 h were obtained. The average duration of an episode is 24 min. The quality of the obtained speech data is reflected in Table 1, which shows that the methodology used indeed yields a fairly high percentage of clean recordings.

Table 1. Proportions of audio recordings of various quality levels in the ESC Corpus.

Quality Levels of Audio Recordings	Percentage
Excellent (1)	58.52
Good (2)	25.40
Satisfactory (3)	9.76
Poor (4)	6.32

Out of 70 informants, 62 (6 men and 56 women) belong to Student/Youth Group, 4—to Young Adult Group (1 man and 3 women), and 4 are older informants (women). The Student/Youth group includes students from 10 universities in Saint Petersburg and Moscow, with the majority of informants (79%) being students from the HSE University in Saint Petersburg, where the corpus is being created.

The gender imbalance is due to the fact that the recordings began in a student philological environment, known for its gender imbalance. It is expected that further additions to the corpus will even out this and other social disparities. The average age of student informants is 20.5 years, most of informants are Russians, with representatives from Kazakhstan, Belarus, and Kyrgyzstan. Half of the informants are from Saint Petersburg,

speech on the main annotation levels required approximately one hour of expert work. After this, the transcript was checked at least twice by other experts, who made their corrections to the initial transcription. This labor-intensive approach to converting audio recordings into text explains why, to date, only a small portion (no more than 25%) of the extensive ORD collection has textual transcriptions, while the majority of the recordings are still awaiting transcription.

and almost all identify Russian as their native language, except for one informant from Belarus.

Results of psychological testing and the passive vocabulary test have been obtained for 70% of the informants. The vocabulary test results range from 68,000 to 106,000 words, with an average of 84,326 words.

The 933 macro episodes prepared for the ESC corpus to date are distributed across the age sub-corpora as follows: the absolute majority of the speech material pertains to student communication, with a small amount of recordings from young adults and middle-aged adults (see Table 2).

Table 2. The number of macro episodes obtained from respondents of different ages.

Youth/Students	Young Adults	Middle-Aged Adults
868 (93.04%)	44 (4.71%)	21 (2.25%)

The statistics presented below pertain to the block of student speech. The total duration of student conversations is 356 h. These conversations were recorded in various locations.

Almost half of the student communication (46.43%) occurs at their primary residence—at home or in a dormitory. Students also frequently communicate outdoors, at their university, visiting friends or relatives, as well as in cafés, restaurants, and other dining establishments. Some students work, so about 4% of the communication occurs in a work office, and a noticeable percentage of conversations take place in public places (see Table 3).

Table 3. Distribution of student communication episodes by location.

Rank	Location	Percentage	Rank	Location	Percentage
1	Home / Dormitory	46.43	7	Public place	4.15
2	Outdoors	11.18	8	Transportation	2.42
3	College / university	10.60	9	Shops	2.30
4	Visiting friends / relatives	7.95	10	Service center	2.19
5	Café / restaurant	7.37	11	Medical center	0.22
6	Office	4.38	12	Unidentified places	0.81

Regarding the distribution of communicative situations encountered by students, the majority (approximately 78.57%) relate to causal or domesic communication, similar to the recordings in the ORD corpus. Communication between colleagues (for working students), educational communication, and client-service communication constitute similar proportions, ranging from 5.3% to 6.8% (see Table 4).

Table 4. Distribution of student communication episodes by communication type.

Rank	Type of Communication	Percentage
1	Casual / Domestic	78.57
2	Professional	6.80
3	Educational	6.45
4	Client-Service	5.30
5	Public	2.88

The variety of places and types of communication also contributes to the diversity of social roles that students assume throughout the day. The processed episodes of the ESC corpus are distributed across social roles as follows (see Table 5). Most often, our student respondents communicate with friends (social role *Friend*), with university professors or classmates (*Student*), with romantic partners (*Girlfriend*, *Boyfriend*), with parents (*Daughter*, *Son*), and with work colleagues (*Colleague*). It is also worth noting some rather unusual communicative roles such as talking to oneself (*Self*) and communicating with their pets (*Pet Owner*).

Table 5. Distribution of students' social roles.

Rank	Location	Percentage	Rank	Location	Percentage
1	Friend	46.43	7	Service Staff	4.15
2	Student	11.18	8	Client	2.42
3	Girlfriend	10.60	9	Self	2.30
4	Daughter / Son	7.95	10	Granddaughter / -son	2.18
5	Colleague	7.37	11	Pet Owner	0.23
6	Boyfriend	4.38	12	Others	0.81

Thus, the prepared section of the corpus demonstrates the diversity of speech domains in everyday student communication. It can serve as a basis not only for linguistic analysis of the speech of contemporary students but also, more broadly, as a foundation for research in speech communication, social, psychological, and anthropological studies.

4 Conclusion

The article describes the methodology for creating a new speech resource—the ESC corpus of everyday speech of students and young people, collected in natural communication environments. Since the recordings for the new corpus have only just begun and will continue, the obtained statistics should be considered preliminary.

The variety of speech material and communicative situations that informants encounter while recording their "speech days", as well as the social roles they assume during these recordings, provide a unique opportunity to use the resulting speech material to study the vocabulary and pragmatics of students in a major Russian city. The planned expansion of the corpus to include speakers from various professional fields, different regions of the country, and not only students but also adults, will enhance the representativeness of the corpus and the reliability of the theoretical and practical works based on it. It also seems advisable to actively involve crowdsourced recordings and increase the number of recordings available for public use. One of the goals of the created resource could be the development of a national corpus of recordings of everyday spoken Russian.

The creation of the corpus will allow for the acquisition of generalized statistical characteristics of the language of contemporary students at various linguistic levels, from phonetics to pragmatics. Comparing the new recordings with the speech material of the ORD corpus, obtained 7 – 15 years ago, will enable the construction of formal models reflecting changes in linguistic characteristics over time. The comparison of the speech material from the ESC and ORD corpora will reveal other possible aspects of the dynamics of speech behavior.

Acknowledgments. This publication was prepared as a result of research conducted for the project "Text as Big Data: Methods and Models for Working with Large Textual Data", carried out within the framework of the Fundamental Research Program of the National Research University Higher School of Economics in 2024. The authors express their gratitude to all the volunteers who participated in the recording of the corpus, as well as to the students of the Philology educational program at the National Research University Higher School of Economics in Saint Petersburg, who actively contributed to the development of the project's methodology and the organization of the recordings and their processing. Special thanks are extended to NTR (https://ntr.ai/) and its director Nick Mikhaylovskiy personally for providing consultations and assistance in speech signal processing.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Asinovsky, A., Bogdanova, N., Rusakova, M., Ryko, A., Stepanova, S., Sherstinova, T.: The ORD speech corpus of Russian everyday communication "One Speaker's Day": creation principles and annotation. In: Matoušek, V., Mautner, P. (eds.) Text, Speech and Dialogue. TSD 2009. Lecture Notes in Computer Science, vol. 5729, pp. 250–257. Springer, Berlin, Heidelberg (2009). https://doi.org/10.1007/978-3-642-04208-9_36
2. ORD Homepage, https://ord.spbu.ru/. Accessed 25 July 2024
3. Reference Guide for the British National Corpus, http://www.natcorp.ox.ac.uk/docs/URG.xml. Accessed 26 July 2024
4. Campbell, N.: Speech & expression; the value of a longitudinal corpus. In: Lino, M.T., Xavier, M.F., Ferreira, F., Costa, R., Silva, R. (eds.). Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, pp. 183–186 (2004)

5. Bogdanova-Beglarian, N., et al.: Sociolinguistic extension of the ORD corpus of Russian everyday speech. In: Ronzhin, A., Potapova, R., Németh, G. (eds.) Speech and Computer. SPECOM 2016, Lecture Notes in Computer Science, vol. 9811, pp. 659–666. Springer, Cham (2016)
6. Akinshina, E., Sherstinova, T.: Thematic diversity of everyday Russian discourse: a case study based on the ORD corpus. In: Mahadeva Prasanna et al. (eds.), Specom 2022, LNCS 13721, Springer Nature, pp. 1–9 (2022)
7. *Russkii iazyk povsednevnovo obshcheniia: osobennosti funktsionirovaniia v raznykh sotsial'nykh gruppakh* (Russian Everyday Communication Language: Features of Functioning in Different Social Groups) / N.V. Bogdanova-Beglarian (ed). SPb: Laika, 2016. 244 p.
8. Fujishima, Y., Yamada, N., Tsuji, H.: Construction of short form of five factor personality questionnaire. Jpn. J. Pers. **13**(2), 231-241 (2004)
9. Perkov, M.A.: Online test "Big Five. Five-factor personality model", https://testometrika.com/psychodiagnostics/five-factor-personality-test-ipip-neo-pi-r/. Accessed 25 July 2024
10. Khromov, A.B.: Piatifaktornyi oprosnik lichnosti (Five-Factor Personality Questionnaire), p. 23. Publishing House of Kurgan State University, Kurgan (2000)
11. MyVocab Homepage, https://www.myvocab.info. Accessed 25 July 2024
12. Read, J.: Assessing Vocabulary. Cambridge University Press, p. 294 (2000)
13. Golovin, G.V.: Izmerenie passivnogo slovarnogo zapasa russkogo yazyka (Measuring the Passive Vocabulary of the Russian Language). In: Socio-i psikholingvisticheskie issledovaniya (Socio- and Psycholinguistic Studies), 2015. No. 3. pp. 148–159
14. Sherstinova, T.: Macro episodes of Russian everyday oral communication: towards pragmatic annotation of the ORD speech corpus. In: Ronzhin, A. et al. (eds.) SPECOM 2015, Lecture Notes in Artificial Intelligence, LNAI, vol. 9319, pp. 268–276 (2015)
15. ELAN Homepage, http://tla.mpi.nl/tools/tla-tools/elan. Accessed 25 July 2024
16. NTR Homepage, https://ntr.ai. Accessed 25 July 2024
17. Radford, A., Kim, J.W., Xu T., Brockman, G., McLeavey, C., Sutskever, I.: Robust Speech Recognition via Large-Scale Weak Supervision, 2022. https://cdn.openai.com/papers/whisper.pdf. Accessed 25 July 2024
18. OpenAI Homepage, https://openai.com. Accessed 25 July 2024
19. Kolpashchikova, E.O.: Pisatel' Robin Dranattagor: aprobatsiya modeli Whisper na russkoyazychnoi zvuchashchei rechi (Writer Robin Dranattagor: Testing the Whisper Model on Russian-Language Spoken Speech). In: Socio- i psikholingvisticheskie issledovaniya (Socio- and Psycholinguistic Studies), Issue 11, pp. 23–27 (2023)
20. Sherstinova, T., Kolobov, R., Mikhaylovskiy, N.: Everyday conversations: a comparative study of expert transcriptions and ASR outputs at a lexical level, Lecture Notes in Computer Science (LNCS), pp. 43–56 (2023)
21. Sherstinova, T., Mikhaylovskiy, N., Kolpashchikova, E., Kruglikova, V.: Bridging gaps in Russian language processing: AI and everyday conversations. In: 35th Conference of Open Innovations Association (FRUCT), April 24–26, Tampere, Finland. pp. 253–258 (2024)

OpenAV: Bilingual Dataset for Audio-Visual Voice Control of a Computer for Hand Disabled People

Denis Ivanko(✉), Dmitry Ryumin, Alexandr Axyonov, Alexey Kashevnik, and Alexey Karpov

St. Petersburg Federal Research Center of the Russian Academy of Sciences, St. Petersburg 199178, Russia
denis.ivanko11@gmail.com, {ryumin.d,axyonov.a,alexey, karpov}@iias.spb.su

Abstract. In recent years, audio-visual speech recognition (AVSR) assistance systems have gained increasing attention from researchers as an important part of human-computer interaction (HCI). The objective of this paper is to further advance the development of assistive technologies in the AVSR field by introducing a multi-modal OpenAV dataset, intended for state-of-the-art neural network model training. The OpenAV is designed to train AVSR models for assistance to persons without hands or with disabilities of their hands or arms in HCI. The dataset could also be useful for ordinary users at hands-free contactless HCI. The dataset currently includes the recordings in two languages (English and Russian) of 15 speakers with a minimum of 10 recording sessions for each. Along with this we provide a detailed description of the dataset and its collection pipeline. In addition, we evaluate state-of-the-art audio-visual (AV) speech recognition approach and present a baseline recognition results. We also describe the recording methodology, release the recording software to public, as well as open the access to the dataset https://smil-spcras.github.io/OpenAV-dataset/.

Keywords: Audio-Visual Dataset · Human-Computer Interaction · Automatic Speech Recognition · Data Collection · Machine Learning

1 Introduction

Nowadays, the creation of a multi-modal assistance system that allows to control a computer without a traditional control device but using audio-visual speech input for giving the control commands is of great interest as well as a big challenge. Many people are unable to operate a personal computer by a standard keyboard, touchpad, or computer mouse because of disabilities of their hands or arms. In this paper we present an audio-visual OpenAV dataset, intended for AVSR model training and creation of a multi-modal assistance system. Using such a system could allow the equal participation of people with disabilities in information society and increase their independence from other people [1].

A. Karpov and V. Delić (Eds.): SPECOM 2024, LNAI 15299, pp. 163–173, 2025.
https://doi.org/10.1007/978-3-031-77961-9_12

In an era where technology strives to enhance accessibility and inclusivity, the development of multi-modal assistance systems that rely on audio-visual speech input represents a significant step forward. Such systems aim to enable users to control computers and other digital devices without relying on traditional input methods like keyboards or mice. This advancement is particularly crucial for individuals with disabilities affecting their hands or arms, who face substantial barriers in using conventional computer interfaces. These systems have the potential to democratize technology access, providing equal opportunities and providing greater independence for users with diverse needs.

Speech recognition is an essential element of HCI, serving as a bridge that connects human-computer communications. However, audio-only speech recognition systems still cannot demonstrate satisfactory accuracy in acoustically noisy environments [2]. To address this problem visual speech signals that remain unaffected by acoustic noise are used to provide supplementary information and improve overall recognition results.

The fundamental challenge in creating effective AVSR systems lies in integrating and processing both auditory and visual data to improve speech recognition accuracy. While audio-based speech recognition systems have made remarkable progress, they often struggle in environments with high levels of background noise or when the speaker's speech is not clear. By incorporating visual information, such as lip movements, AVSR systems can mitigate these issues, offering a more robust solution. Visual cues provide context that helps disambiguate spoken words and phonemes, enhancing the system's ability to accurately interpret and respond to commands even in noisy conditions.

To solve the above-mentioned challenges, we provide the following novel contributions: (1) we present a new audio-visual OpenAV speech dataset, designed to train AVSR models and create a multi-modal assistance system for persons with hands disabilities. OpenAV consists of recordings of 15 speakers uttering the script of 26 voice control commands in Russian with at least 10 recording sessions for each); (2) we provide a detailed description of the recording pipeline and framework, as well as release the web-based recording service to the public use; (3) we evaluate state-of-the-art audio-visual speech recognition approach and present a baseline recognition results for the OpenAV dataset.

This paper is structured as follows: following the Introduction Sect. 2 provides a brief overview of research related to AVSR assistance systems and datasets; Sect. 3 reveals the recording methodology; Sect. 4 presents the OpenAV dataset and its main parameters; Sect. 5 provides experimental results on AVSR task; conclusions and future research directions are given in Sect. 6.

2 Related Works

Recent advancements in deep learning and neural network architectures have significantly contributed to the progress in audio-visual speech recognition (AVSR) [3]. These systems leverage both audio and visual data to enhance the robustness and accuracy of speech recognition, particularly in challenging environments where audio data alone might be insufficient. The integration of visual cues, such as lip movements, provides additional context that can help disambiguate phonemes that are otherwise difficult to

distinguish due to noise interference. This multimodal approach not only improves performance in noisy settings but also makes the technology more accessible and practical for real-world applications.

The design and development of modern AVSR systems is a data-driven process and is inevitably affected by the amount and quality of available data. In order to develop real-world audio-visual assistance systems, high-quality training and testing datasets are vital.During the last decade AVSR has been a popular research direction for the speech recognition community. Acquisition of multimodal or audio-visual speech dataset is a challenging task due to many reasons, such as subjects, illumination, noise, head-pose, vocabulary, resolution, etc. The lack of suitable datasets has been one of the major obstacles to progress in the AVSR field for a long time [4].

The field of audio-visual speech recognition has made remarkable progress in recent years, driven by advancements in deep learning and the availability of high-quality datasets. The ongoing development of innovative architectures and fusion techniques promises to further enhance the robustness and accuracy of AVSR systems, making them increasingly viable for practical applications in diverse environments.

The general idea of AVSR is to recognize speech based on the processing of both audio and video signals simultaneously. Several deep neural network architectures, training strategies and audio-visual fusion techniques have been proposed for AVSR [5].

The selection of a proper AV speech dataset is eventually a key task in building speech recognition systems due to state-of-the-art solutions are usually based on data-driven machine learning methodology [6]. The database must meet the requirements of deep learning in terms of the number of speakers, vocabulary size, number of repetitions, etc. Compliance with all the necessary parameters is quite challenging [7].

The researchers in the works [8] and [9] provide a comprehensive analysis of existing audio-visual speech databases. In this paper we refrain from repeating existing research and refer readers to the aforementioned papers. In general, existing databases can be divided into two groups [10].

The first group includes large-scale publicly available corpora collected from internet sources, such as YouTube, television: LRW [11], LRS-BBC [12], LRS-TED [13], MuAViC [14], MAVD [15], etc. The second group includes much smaller datasets recorded under controlled conditions and specifically designed toward certain tasks, e.g. analysis of frame rate influence [16], of certain language other than English [17], of Lombard effect [18], of speech in a driving condition [19], etc.

The combination of state-of-the-art deep learning methods and audio-visual datasets has been highly successful, achieving significant recognition accuracy results and even surpassing human performance in many tasks [20]. However, according to our analysis there are no publicly available AV datasets intended for speech command recognition of HCI assistance systems which are suitable for state-of-the-art neural network model training. There is no common benchmark dataset, especially for the Russian language. Our goal in releasing the OpenAV dataset is to fill in this scientific gap and to provide such a benchmark.

3 Recording Pipeline

The pipline we used for recording the dataset is shown in Fig. 1. The core is a recording session that is supported by our developed web-service https://cais.iias.spb.su/scripts/speech-dataset/. To record the dataset that supports voice control functionality of the computer we proposed to record the following vocabulary (see Table 1).

Recording bilingual datasets for AVSR is crucial for several reasons, particularly in the context of developing more inclusive and versatile technology. As global communication increasingly bridges linguistic and cultural boundaries, the need for AVSR systems that can accurately process and understand multiple languages becomes ever more pertinent. Bilingual datasets serve as a foundational element in building robust AVSR models capable of operating effectively across diverse linguistic environments.

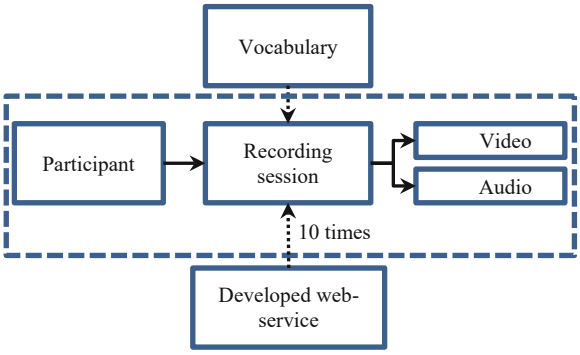


Fig. 1. Dataset recording methodology.

The vocabulary includes basic functions that a user can use to control the computer. We mention that there are a lot of audio-based datasets for English and Russian language that support the following functionality but at the moment the audio-visual dataset is missing.

We used this vocabulary for every recording session. To automate the recording procedure, we built a web-service to collect the OpenAV dataset as shown in Fig. 2. We used the following technologies to build the web-service: Hypertext Markup Language (HTML), Cascading Style Sheets (CSS), and JavaScript.

The development of the web-service for dataset collection was a critical component of our methodology. This web-service not only facilitated the efficient recording and storage of audio-visual data but also provided a user-friendly interface for managing the recording sessions. Users could easily start, pause, and stop recordings, as well as review and annotate the data in real-time. This streamlined process significantly reduced the time and effort required for dataset collection, making it feasible to scale up the number of speakers and commands recorded.

The web-service allows participants to perform a video recording. It shows the phrase from the vocabulary that the participant should pronounce. The web-service use connected camera to record the audio-visual data. We provide all the instructions for recording audio and video for our dataset. In our opinion this demonstrates the easiest and

Table 1. Proposed vocabulary for OpenAV dataset recording.

	Russian	English	Class		Russian	English	Class
1	Левая	Left	Mouse manipulator commands	14	Новый	New	Grafical User Interface commands
2	Правая	Right		15	Открыть	Open	
3	Нажать левую	Left down		16	Сохранить	Save	
4	Отпустить левую	Left up		17	Вырезать	Cut	
5	Нажать правую	Right down		18	Копировать	Copy	
6	Отпустить правую	Right up		19	Закрыть	Close	
7	Двойной клик	Double click		20	Вставить	Paste	
8	Вверх	Scroll up		21	Печать	Print	
9	Вниз	Scroll down		22	Вперед	Next	
10	Ввод	Enter	Keyboard button commands	23	Назад	Previous	
11	Отменить	Escape		24	Выделить все	Select all	
12	Удалить	Delete		25	Пуск	Start	
13	Выключить	Shut down		26	Калибровка	Calibration	

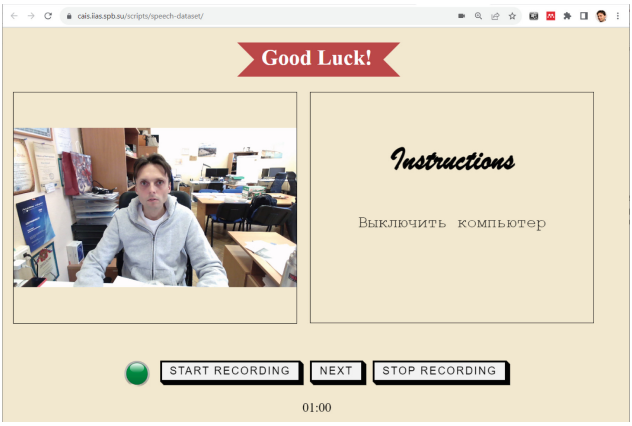


Fig. 2. Developed web-service interface.

most convenient way for collecting data, since it can be accessed from all devices and all operating systems. Participants only need an internet connection and a web browser. We recorded FullHD video (1280x720) of 30 fps with 48 kHz for audio.

We perform dataset recording in conditions that correspond to the environment where the assistive command recognition system will be used. In our case it is an office environment that has mostly good lighting condition as well as an environment without heavy acoustic noise.

4 OpenAV Dataset Description

The dataset currently includes recordings of 15 speakers in an office environment (Fig. 3). It is worth noting that the dataset can easily be expanded with additional speakers or a new dictionary, considering the developed recording methodology. Each speaker utters 26 voice control commands in Russian and English languages to operate their personal computer. The dictionary categorizes commands as "Mouse manipulator commands", "Keyboard button commands", and "Graphical User Interface commands". The speakers were positioned in a controlled office environment to minimize background noise and distractions. These commands were carefully selected to cover a wide range of typical interactions a user might have with a computer, from basic cursor movements and clicks to complex sequences involving multiple keyboard inputs and interface navigation.

Fig. 3. Snapshots of the speakers during recording session.

The main parametric characteristics of the recorded dataset are depicted in Fig. 4. In addition to the basic commands, the dataset includes variations in speech delivery, such as different speaking rates, intonations, and accents. These variations are essential for creating robust AVSR systems capable of handling real-world variability in speech patterns. By incorporating such diversity, the OpenAV dataset aims to cover a wide range of potential use cases.

A critical feature of the OpenAV dataset is the diversity of its speakers. The dataset includes recordings from 15 different speakers, representing various genders, ages, and speech characteristics. This diversity is essential for training AVSR models that can generalize well across different users. Each speaker was recorded in a consistent office environment to control for external variables, but individual differences in speech patterns, accents, and facial movements provide a rich source of variability for the models to learn from. This diversity helps ensure that the resulting AVSR systems are robust

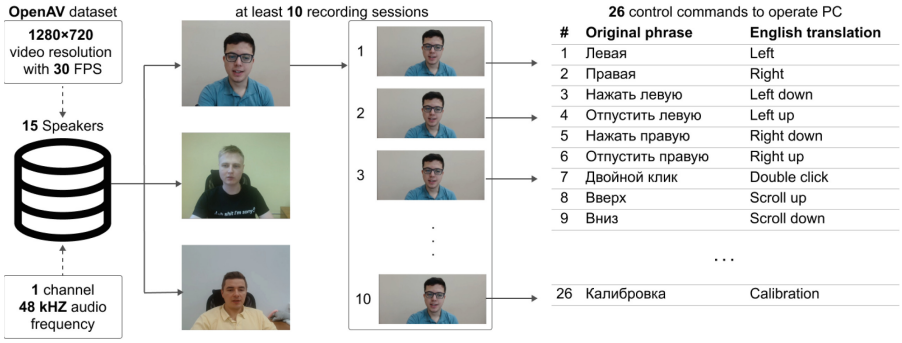


Fig. 4. Main characteristics of the OpenAV dataset.

and capable of performing well in real-world scenarios where speaker characteristics can vary widely.

One of the primary benefits of bilingual datasets is their ability to improve the generalization and adaptability of AVSR systems. Multilingual capabilities allow these systems to be used by a broader audience, catering to speakers of different languages without requiring separate models for each language. This not only enhances user experience by making technology more accessible but also reduces the development effort and resources required to build and maintain multiple language-specific models. By training on bilingual data, AVSR systems can learn to recognize and interpret commands in various languages, thus accommodating users in multilingual regions or those who frequently switch between languages.

Bilingual datasets contribute to the robustness of AVSR systems by exposing them to a wider range of phonetic variations, accents, and linguistic nuances. Language-specific phonetic characteristics, such as the different articulatory patterns and phoneme inventories, can affect speech recognition performance. Training on bilingual data helps the model become more adept at distinguishing between these variations and enhances its ability to handle cross-linguistic phonetic overlaps. This, in turn, improves the overall accuracy and reliability of the system, making it more effective in real-world scenarios where users may speak with diverse accents or in noisy environments.

We also explored the potential for expanding the dataset to include more complex interactions and additional languages. Future work could involve integrating more diverse linguistic elements and contextual scenarios, such as additional languages, conversational speech and multi-speaker interactions. This would further broaden the applicability of the dataset and enhance the AVSR models' ability to generalize across different contexts.

5 Evaluation Experiments

In this section, we propose and evaluate the AVSR model architecture and recognition pipeline. We assess the proposed NN architecture using the collected OpenAV dataset. It has been splitted into the Train, Validation and Test subsets of 70%, 20% and 10% number

of recordings, respectively. The method utilises two open-source libraries: MediaPipe's Face Mesh [21] is utilized for video preprocessing, while Librosa [22] is applied for audio preprocessing.

The dataset boasts high-definition video and high-quality audio recordings. The combination of high-quality audio and visual data is crucial for training effective AVSR systems, as it allows models to accurately learn the relationship between spoken words and corresponding visual cues.

Initially, lip region images are acquired using the MediaPipe Face Mesh algorithm. To account for differences in lip shape, facial proportions, and articulation, all images are normalized to a size of 88 × 88 × 3 by filling in missing pixels with average values. Since each video comprises of 30 frames per second, the sequence length is 30 images. Using Librosa, a logarithmic Mel spectrogram with 128 Mel filters is created from the audio signal, using a short-time Fourier transform window size of 2048 and a step of 128. This results in an image with dimensions of 128 × 251 × 3. The dimension of 251 is a consequence of the 2-s duration of the window, the sampling rate of 16K Hz, and the use of a Fourier window step of 128. Afterward, min-max normalization is applied to these images.

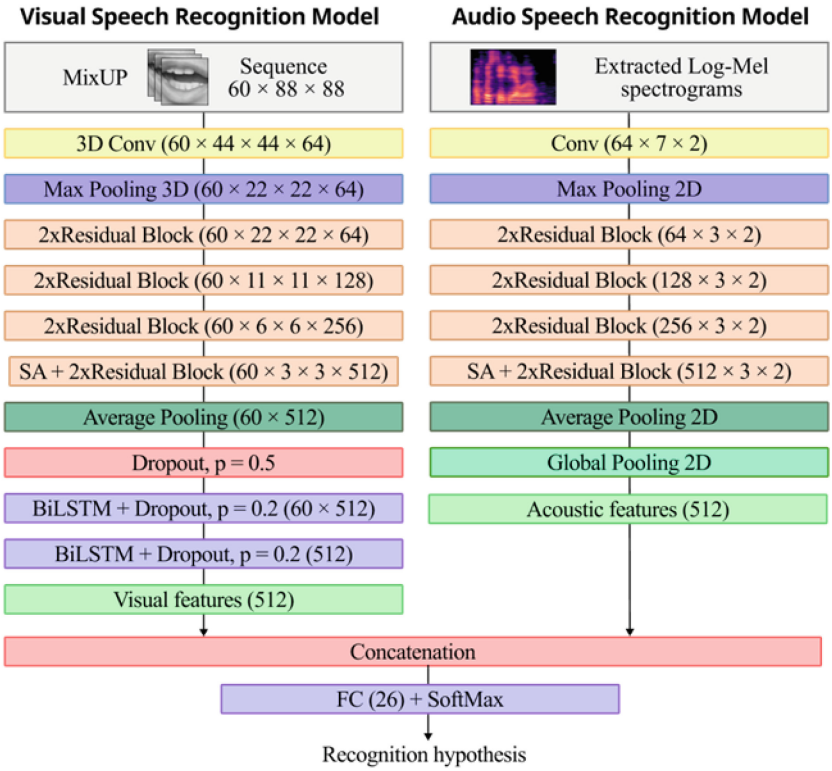


Fig. 5. Audio-visual neural network model architecture. Visual: 3DResNet18 + SA + BiLSTM; Audio: 2DResNet18 + SA; Fusion: Model-level.

The general neural network model architecture is shown in Fig. 5. The input data for the audio-visual model comprises of two components for handling visual and audio information, both incorporating the ResNet-18 architecture.

The trained neural network model is fed with segmented image data consisting of 60 frames, which is then spatially and temporally convoluted by the 3D convolution layer. The MaxPooling 3D subsampling layer is utilised to extract the most significant features from the sequences. The modified residual blocks of the ResNet18 model extract $60 \times 3 \times 3 \times 512$ feature maps from each frame of the input sequence. The Squeeze-and-Attention (SA) mechanism [23] is included in the final residual block. The Average Pooling subsampling layer converts the obtained feature maps into 60×512 one-dimensional vectors, which are fed into the next two bidirectional Long Short-Term Memory (LSTM) layers. The first layer is sequence-to-sequence with input and output 60 feature vectors, the second layer is sequence-to-one with output a 512-dimensional feature vector.

The acoustic neural network model architecture inputs Log-Mel spectrogram images extracted from the audio signal. The trained neural network model convolves input Log-Mel spectrogram images with a spatio-temporal convolution layer. In order to extract the most significant features from the Log-Mel spectrogram, the dimensionality of the input vector is reduced by means of a MaxPooling 2D subsampling layer. The visual model generates a 512-sized feature vector, as does the audio model. These feature vectors are then concatenated and passed onto the final fully connected neural network for prediction. Both models share the same training parameters, including learning rate, schedule, optimizer, and batch size, which allows concurrent learning from acoustic and visual data. This method, called model-level fusion, mimics human cognitive processing by combining audio and visual information.

The choice of model structures in this approach was determined through a series of experiments. The fusion of both audio and visual modalities results in an accuracy of 91.54%, achieved through a model-level fusion approach.

6 Conclusion

In this paper, we have created a bilingual multi-speaker audio-visual dataset OpenAV, designed for state-of-the-art neural network model training and intended for building AVSR assistance systems for people with hands disabilities. The dataset could also be useful for ordinary users at hands-free contactless HCI. The dataset currently includes the recordings of 15 speakers with a minimum of 10 recording sessions for each in two languages English and Russian. Along with this we have provided a detailed description of the dataset and its collection pipeline.

In addition, we have evaluated state-of-the-art audio-visual (AV) speech recognition approach and have presented a baseline recognition results. The fusion of both audio and visual modalities results in an accuracy of 91.54%, achieved through a model-level fusion approach for both languages. This in terms of recognition accuracy is comparable to the state-of-the-art results achieved for other AV corpora.

Developing effective AVSR systems involves creating large, diverse, and high-quality datasets that encompass various speech scenarios and environmental conditions.

The OpenAV dataset presented in this paper is designed to address these requirements by offering a comprehensive collection of synchronized audio and visual speech recordings in two languages, English and Russian. This dataset supports the training of AVSR models capable of recognizing speech with high accuracy across different speakers. By utilizing the OpenAV dataset, researchers can develop more robust models that push the boundaries of current speech recognition technologies, paving the way for innovative applications in assistive technologies and beyond.

One of the most significant advantages of the OpenAV dataset is its expandability. The recording methodology and technological framework are designed to accommodate additional speakers and commands seamlessly. This flexibility means that the dataset can be continuously updated and expanded, allowing for the inclusion of new languages, additional command sets, and more diverse speaker demographics. Such expandability ensures that the dataset remains relevant and useful as the field of AVSR evolves and as new requirements emerge.

In addition, recording bilingual datasets is essential for advancing AVSR technology, making it more inclusive, versatile, and robust. Such datasets enable systems to cater to a wider audience, handle various linguistic features and enhance performance in multilingual contexts. As the demand for effective and accessible technology continues to grow, bilingual or multi-lingual datasets will play a crucial role in ensuring that AVSR systems can meet the diverse needs of users worldwide.

We have also released the recording software to the public, as well as opened the access to the dataset for non-commercial use https://smil-spcras.github.io/OpenAV-dataset/.

Acknowledgments. Sections 1, 2, 4, 5 and 6 are financially supported by the Russian Science Foundation (project No. 23–71-01056), and Section 3 by the State Research № FFZF-2022–0005.

References

1. Karpov, A., Ronzhin, A., Kipyatkova, I.: An assistive bi-modal user interface integrating multi-channel speech recognition and computer vision. In: Human-Computer Interaction. Interaction Techniques and Environments: 14th International Conference, HCI International 2011, Orlando, FL, USA, July 9–14, 2011, Proceedings, Part II 14, pp. 454–463 (2011)
2. Wang, J., et. al.: Restoring speaking lips from occlusion for audio-visual speech recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 17, pp. 19144–19152 (2024)
3. Ma, P., Petridis, S., Pantic, M.: End-to-end audio-visual speech recognition with conformers. In: ICASSP 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7613–7617 (2021)
4. Ryumin, D., et. al.: Audio–visual speech recognition based on regulated transformer and spatio–temporal fusion strategy for driver assistive systems. In: Expert Systems with Applications, vol. 252, p. 124159 (2024)
5. Shi, B., Hsu, W. N., Mohamed, A.: Robust self-supervised audio-visual speech recognition. In: Interspeech 2022 (2022)
6. Burchi, M., et. al.: Multilingual audio-visual speech recognition with hybrid CTC/RNN-T fast conformer. In: ICASSP 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 10211–10215 (2024)

7. Chen, C., et. al.: Leveraging modality-specific representations for audio-visual speech recognition via reinforcement learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 11, pp. 12607–12615 (2023)
8. Fernandez-Lopez, A., Sukno, F. M.: Survey on automatic lip-reading in the era of deep learning. In: Image and Vision Computing, vol. 78, pp. 53–72 (2018)
9. Ivanko, D., Ryumin, D., Karpov, A.: A review of recent advances on deep learning methods for audio-visual speech recognition. Mathematics, **11**(12), 2665 (2023)
10. Wang, X., et. al.: CATNet: cross-modal fusion for audio–visual speech recognition. Pattern Recogn. Lett. **178**, 216–222 (2024)
11. Chung, J.S., Zisserman, A.: Lip reading in the wild. In: Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20–24, 2016, Revised Selected Papers, Part II 13, pp. 87–103 (2017)
12. Chung, J.S., et. al.: Lip reading sentences in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6447–6456 (2018)
13. Afouras, T., Chung, J.S., Zisserman, A.: LRS3-TED: a large-scale dataset for visual speech recognition. In: arXiv preprint arXiv:1809.00496 (2018)
14. Anwar, M., et. al.: MuAViC: a multilingual audio-visual corpus for robust speech recognition and robust speech-to-text translation. In: Interspeech 2023, pp. 4064–4068 (2023)
15. Wang, J., et. al.: MAVD: the first open large-scale mandarin audio-visual dataset with depth information. In: Interspeech 2023, pp. 2112–2117 (2023)
16. Verkhodanova, V., et. al.: HAVRUS corpus: high-speed recordings of audio-visual Russian speech. In: Speech and Computer: 18th International Conference, SPECOM 2016, Budapest, Hungary, August 23–27, 2016, Proceedings 18, pp. 338–345 (2016)
17. Zhao, Y., Xu, R., Song, M.: A cascade sequence-to-sequence model for Chinese Mandarin lip reading. In: Proceedings of the ACM Multimedia Asia, pp. 1–6 (2019)
18. Alghamdi, N., et. al.: A corpus of audio-visual Lombard speech with frontal and profile views. J. Acoust. Soc. Am. **143**(6), 523–529 (2018)
19. Ivanko, D., et. al.: RUSAVIC corpus: Russian audio-visual speech in cars. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC), pp. 1555–1559 (2022)
20. Ryumin, D., Ivanko, D., Ryumina, E.: Audio-visual speech and gesture recognition by sensors of mobile devices. Sensors **23**(4), 2284 (2023)
21. Lugaresi, C., et. al.: Mediapipe: a framework for perceiving and processing reality. In: Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) (2019)
22. McFee, B., et. al.: librosa: audio and music signal analysis in Python. In: Proceedings of the 14th Python in Science Conference, vol. 8, pp. 18–25 (2015)
23. Zhong, Z., et. al.: Squeeze-and-attention networks for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13065–13074 (2020)

Bulgarian Speech Resources in the CHILDES System

Velka Popova and Dimitar Popov(✉)

Konstantin Preslavsky University of Shumen, Universitetska str. 115, 9700 Shumen, Bulgaria
labling@shu.bg
http://labling.fhn-shu.com/home.htm

Abstract. The main aim of the article is to present the Bulgarian speech resources in the CHILDES automated computerized language data exchange system. The resources are organized into a corpus labeled Bulgarian LabLing Corpus (https://childes.talkbank.org/access/Slavic/Bulgarian/LabLing.html). This database is the most recent contribution to the CHILDES Slavic collection, the first version of which was published in 2020. It covers longitudinal data from 5 young children and the picture-based narratives of 121 preschoolers. The focus in the first Bulgarian electronic corpus with children's speech is on the possibilities for the study of child-adult speech interaction, which is extremely important for the adequate modeling of language ontogenesis.

Keywords: Child speech resources · CHILDES · Bulgarian LabLing corpus

1 Introduction

The necessity of studying child speech is undeniable for modern science, the irrefutable proof of which is the wide application of ontogenetic data as evidence in solving various problems of linguistics. Indicative of this is what Stefan Mladenov said back in 1930s: "In general, the development of child speech provides material for illuminating and even for resolving all major and minor issues in linguistic history, not only in the fields of phonetics and morphology, but also of etymology, word formation, lexicology and syntax ... Those who know nothing of the development of child speech know nothing of the linguistic production of adults" [1].

The importance of ontogenetic data for linguistics alone does not explain the need to create a corpus of child speech. Moreover, the study of this exotic and peculiar phenomenon has its own long and rich biography, in which, however, the problem of the reliability of both the empirical material and the methods for its collection, systematization and processing, has always remained open. At the same time, with the imposition of the anthropocentric paradigm in modern humanities, significant changes are taking place in the overall concept of the study of children's speech, in which the focus is no longer on the isolated study of its units and features, but on the speaking child instead. From this point of view, the turn of scientists to the holistic paradigm, which supports the thesis of interweaving of common linguistic and cognitive principles and rules, is

A. Karpov and V. Delić (Eds.): SPECOM 2024, LNAI 15299, pp. 174–186, 2024.
https://doi.org/10.1007/978-3-031-77961-9_13

completely understandable. This, in turn, shifts the interest from the traditional linguistic explanation of linguistic rules and principles, to the search for relevant answers from the perspective of cognitive linguistics. The study of the nature of the linguistic cognitive system and the access to its processor is now highlighted as a key research problem.

This raises the question of the empirical assurance of the cognitive holistic paradigm. The possible solutions can be found in the field of modern corpus linguistics, since it is precisely in this field that databases are produced, solid not only in terms of their volume, but also in terms of the possibilities for complex study of speech. Therefore, corpus studies carried out in the light of the holistic tradition in linguistics, have dominated as a research paradigm in the last decades. And thanks to the "achievements of modern information technologies, and the created multi-media program packages designed to achieve various goals in the branches of applied linguistics, it became possible for the modern researcher to illustrate and visualize the results in a complex and comprehensive manner. The accumulated research in the new multimodal linguistics, representing one of the directions for the simultaneous multidimensional presentation of linguistic facts and phenomena within the framework of holistic linguistics, also contribute to it [2].

In the light of what has been said, the present article is dedicated to the corpus approach in the study of children's speech, whose advantages are presented through the example of some modern well-functioning research platforms such as CHILDES (https://childes.talkbank.org/) and TalkBank (https://www.talkbank.org/). More specifically, the first electronic corpus of Bulgarian children's speech (Bulgarian LabLing Corpus) is in focus. At the same time, an attempt is made to highlight the wide possibilities of both the chosen format for presenting the data in it, implemented in terms of the interactive platforms TalkBank and CHILDES, and the corpus perspective in modern linguistics, thanks to which there is an optimal environment for creating objective models of language, and for limiting the danger of the emergence of new myths in modern scientific research.

In this line of thought, the question arises quite naturally as to whether it is justified to invest effort and time in such a laborious undertaking when there are sufficiently good alternatives in the research tradition. Could the creation of a corpus of children's speech not be interpreted as some fad? What is wrong with the already known and well-functioning traditional models? In response to these questions, a brief chronological review of existing contributions to the collection and organization of empirical data will be offered, and in this context the place and significance of modern electronic formats of corpus linguistics will be sought.

2 The Corpus Perspective in Child Speech Research

Child speech is an exotic and unique phenomenon that has its own long and rich biography, in which, however, the problem of the reliability of both empirical material and the methods for its collection, systematization and processing has always remained open. In this sense, the present article argues that the corpus perspective could be defined as dominant in the field of language ontogeny studies from the times of Charles Darwin

till the present day. In support of this, one could cite plentiful existing evidence in the ontolinguistic tradition which supports the fact that the accumulation of empirical data has always been, is, and will be dominant. This evidence will be presented briefly in the present paper, in order to highlight the specifics in the evolution of the development of the speech corpora, which are extrapolated in an adequate format for the relevant research period. In this regard, several successive stages could be distinguished in the history of child speech studies:

- The stage of the so-called "baby biographies" from mid-19th to the mid-20th century, which are associated with the practice of keeping diaries of the chronological development of child speech, which are mostly descriptive (as the detailed notes on the speech development of their children made by philosophers, naturalists, linguists, psychologists, among whom are Hippolyte Taine, Charles Darwin, D. Tiedeman, V. Leopold, Gregoire, Jan Baudouin de Courtenay, A. N. Gvozdev, I. Georgov.).
- The stage of the appearance of the first cross-sectional studies in the 1930s, in which samples of many children of the same age could be compared, allowing researchers to apply a variety of statistical methods, to plan and conduct experiments.
- The stage of longitudinal cross-sectional studies, the beginning of which is associated with 1960s, in which a kind of synthesis of the methodological achievements of the two previous stages is observed, insofar as they create an opportunity to overcome the fragmentation and randomness inherent in diaries and of the sample data.
- The stage of creation of computer systems for accumulation and automatic processing of huge amounts of child speech data, allowing researchers to overcome the shortcomings of the previous methods. Despite the usefulness of diaries for the study of the ontogeny of the lexicon, they are not suitable for obtaining reliable quantitative results. Cross-sectional studies, in turn, provide a voluminous database, but they are unable to sufficiently account for the individual cases in language acquisition. Longitudinal studies provide a relatively accurate picture of the individual child, but the laboriousness of data collection and the different methods of transcribing or processing in general make comparison with other longitudinal studies extremely difficult, and the study of just one single corpus could be misleading due to predetermined individual differences in the process of acquisition.

The last stage in the research of linguistic ontogenesis can be defined as qualitatively new, since it is the computer systems created within its scope for the accumulation and automatic processing of huge amounts of child speech data in recent decades create conditions for the successful implementation of large-scale cross-linguistic projects including multiple languages. Moreover, linguistic resources organized as corpora are increasingly used for the purposes of modeling the language and the speech behavior of its speakers, despite the fact that creating and maintaining computer corpora is an extremely labor-intensive and expensive undertaking.

With the advent of modern technologies, a new, more efficient standard for their presentation and processing is required. In this way, it becomes possible to extend the scope of a single corpus to millions of language units, while also optimizing the options for their annotation, unification, standardization and reuse.

Thus, with the development of technical progress over time, card files and diaries are replaced by electronic speech data, and the time-consuming and exhausting work of registering, transcription and statistical processing data is supported by various technical means and software products. The creation of systems such as CHILDES and TalkBank marks the apogee of the evolutionary process summarized here. The typological diversity of the included linguistic data, the uniform transcription format, the CLAN program resources for automatic processing, make this system an extremely useful and convenient platform for conducting research. At the same time, it can be added that it is precisely the optimal empirical possibilities that could guarantee that every single linguistic study has achieved a high degree of objectivity and adequacy of the obtained results, as well as being a solid basis for approbation of the models of language ontogenesis. In the context of this, the choice of the CHILDES platform for the creation of the Bulgarian corpus of child speech data, which is presented in this paper, is completely understandable.

In the next part of the exposition, an attempt will be made to give an idea of the new quality of the research from the last stage with regard to Bulgarian ontolinguistics. Thus, quite naturally, the automatic computer system CHILDES comes into focus as one of the most popular data exchange platforms for child speech especially one of its corpora, the Bulgarian LabLing Corpus. The main goal is to demonstrate that this format of the presented Bulgarian collection of child speech data is not the fruit of a passing linguistic fad, but rather an opportunity to provide reliable empirical material for creating adequate ontogenetic models, as well as for achieving a qualitatively new standard of research.

At the same time, turning to CHILDES is also predetermined by the fact that, on the example of this system, the possibilities of applying corpus linguistics in the study of child-adult speech interaction can be highlighted as much as possible, which could provide answers to many still open questions related to language ontogenesis and the acquisition of the Bulgarian language at the dawn of human life.

3 Some Terminological Clarifications

Before proceeding to the presentation of the Bulgarian CHILDES corpus, it is necessary to specify some terms related to child speech. It should immediately be noted that the term falls into the networks of exceptional pluralism in modern linguistic terminology, and it is associated with three different concepts in the context of the Multilevel Model of Bulgarian Speech [3], in which the trichotomy standard - substandard - nonstandard is embedded [4]. As a result of the specific approbation of the Model when establishing the concepts child speech 1 (the speech of the child in the process of communicating with adults or other children), child speech 2 (the speech of an adult to a child) and child speech 3 (stylized child-manner speech of an adult when communication with adults), they turned out to be related respectively to the substandard in the process of acquisition, to the substandard, and to the nonstandard [5].

As far as the article, as mentioned in the Introduction, is focused on the adult-child interaction, after the above clarifications, in the following exposition, attention will be directed to the two substandard phenomena, namely – child speech 1 and child speech

2. It is their adequate understanding that would lead to the creation of working models of language ontogenesis, insofar as the specific register for communication between adults and children contains some important prerequisites for the successful and rapid acquisition of language at an early age, since it is in it that the target language system is presented in the form of a minimum input program. Thus, in the course of communication with children, adults provide input linguistic data, among which the so-called positive and negative data clearly stand out [6].

In line with these thoughts about the significance of the so-called "Baby talk" for the child's mastery of language, the need to study not only child speech, but also the social variation of language, known as adult-to-child register, naturally emerges. In the last few decades, research contributions have been accumulated, among which longitudinal studies are particularly valuable, thus creating conditions to adequately capture the logic of language development. Undoubtedly, this was also accompanied by searches for suitable formats of speech collections that would give optimal access to both child speech and adult speech to children. That is why in the present paper the optimal system for differentiated extraction of the data of the individual subjects CHILDES was chosen as a focus of attention.

At the same time, CHILDES is an optimally open system in which, in addition to longitudinal data, one can present data from cross-sectional studies, such as collections of narrative samples drawn from multiple children of the same age. It is this variety of data that characterizes the first Bulgarian CHILDES corpus (Bulgarian LabLing Corpus) presented here.

4 Bulgarian Speech Resources on the CHILDES Platform – *Bulgarian LabLing Corpus*

The CHILDES database contains a large amount of information on the acquisition of multiple languages extracted from different language groups, and their number is constantly growing. In the autumn of 2020, a new addition appeared in the Slavic Collection of the CHILDES platform, namely the Bulgarian LabLing Corpus in its pilot version, which was further developed and expanded to its final format in 2023 (see Fig. 1).

Bulgarian LabLing Corpus can be freely accessed at: https://childes.talkbank.org/access/Slavic/Bulgarian/LabLing.html (Fig. 2).

The corpus of Bulgarian child speech presented here is the result of the extended work of researchers from LABLING. Corpus data have been transcribed into the CHILDES unified CHAT format [7], making them comparable to the corpora of other languages on the platform. The CLAN program package allows to make a different type of analysis of the entered dialogues (phonetic, morphological, syntactic) and the comments to them. In this sense, CLAN provides the possibility to automatically obtain the most diverse statistical and substantive results from the transcribed and coded data, such as data about word frequency, lexical diversity and coherence, specific words and forms used in the respective speech session (such as children's language errors as specific deviations from the norm of the respective language: units of the so-called BABY TALK, onomatopoeia, overgeneralizations, children's and family's occasionalisms, etc.)

In the structure of the Bulgarian LabLing Corpus there are two subcorpus, one of which is longitudinal and the other cross-sectional: Corpus A with the longitudinal data of 5 Bulgarian very young girls; Corpus B with 233 narratives of 121 preschool children, which can be defined as cross-sectional, insofar as the subjects are further divided by age into three groups.

Corpus A covers the transcribed longitudinal data obtained from 5 Bulgarian girls – ALE (aged 1; 01.29–2; 04.09), TEF (aged 1; 03.11–2; 05.25), BOG (aged 2; 01.09–2; 04.11), ELI (aged 1; 01.07–1; 07.22), SIM (aged 1; 04.14–3; 01.00). The studied children were born and at the time of the experiment lived in Northeastern Bulgaria. They were recorded in everyday situations (playing, dressing, eating, falling asleep, looking at picture books, etc.) in the process of their daily communication with their loved ones. All persons registered in the database as participants in the dialogues are monolingual

CHILDES Slavic Corpora

Corpus	Age Range	N	Media	Comments
Bulgarian				
LabLing	1-5	5, 50, 71	some audio	5 longitudinal, 50 and 71 narrative
Croatian				
Kovacevic	1;3-2;8 1;10-2;11 0;10-3;2	3	audio	Two girls and a boy learning Croatian in Zagreb
MAIN	5-63	143	audio	MAIN protocol
Czech				
Chromá	1;5-3;5	7	audio	children learning Czech in Prague
Polish				
Szuman	1;5-7;9	10	-	Diary data collected by Szuman and his students and computerized by Magdalena Smoczynska
WeistJarosz	1;7-2;6	3	audio	in PhonBank
Polish-CDS	adults	various	-	child-directed speech from various corpora
Russian				
Protassova	1;6-2;10	1	-	Longitudinal study of a child learning Russian
Tanja	2;5-2;11	1	-	A child learning Russian in a monolingual environment in the United States
Serbian				
SCECL	1;6-4;0	8	audio	Recordings in homes with many people included
Slovenian				
Zagar	5;0	20	-	Arguments patterns in kindergarten children

Fig. 1. The Slavic Collection of the CHILDES platform.

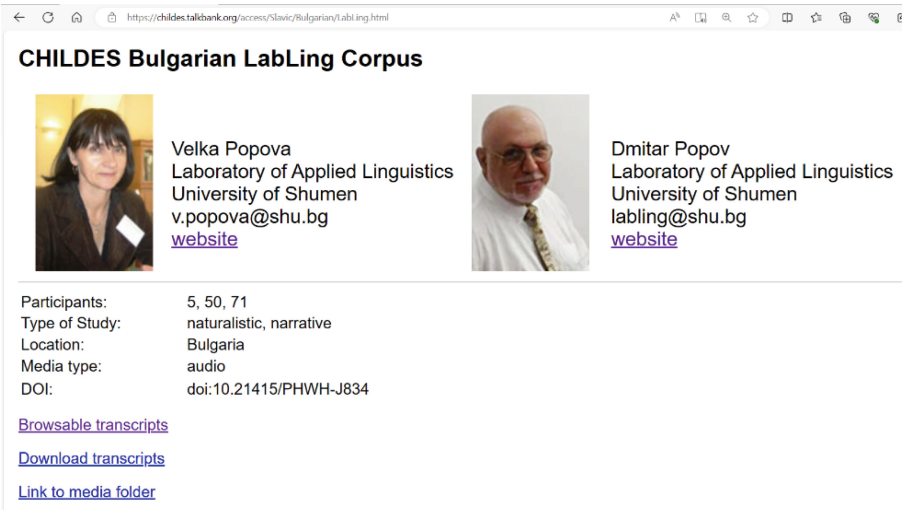


Fig. 2. CHILDES Bulgarian LabLing Corpus.

speakers of the Bulgarian language. Adults from the children's environment have a good level of education (secondary high school and university).

Corpus B contains two segments, namely the Fox-Cat Collection (including 91 children's stories) and the Dog-Birds Collection (142 children's stories), in which the narratives of the 121 children based on the respective picture series are covered. At the same time, the data are structured in view of the reference of the examined persons to one of the three age groups – children aged 3–4 years for the first group, 4–5 years for the second group, 5–6 years for the third group.

The first collection of Corpus B is based on two picture stories: Cat Story [8] and Fox Story [9], each of which consists of 6 black and white drawings without a text. The second narrative collection of Corpus B is based on two colored picture stories – Baby Birds and Dog Story, borrowed from MAIN: The Multilingual Assessment Instrument for Narratives [10].

In the creation of Corpus B, the selection of picture stories and their format (black and white or colored) were not randomly selected. Specially developed materials, which have already been used to create uniform corpora with narratives in many other languages were used, corresponding to the MAIN unified system, designed for the analysis of child narratives [10,11].

The data in both sub-corpora of the corpus with Bulgarian child speech presented here are transcribed in the unified CHAT format of the CHILDES system, which makes them comparable to the corpora of other languages from the platform. CLAN programs allow for a different type of analysis of dialogues and narrativesq as CLAN allows to automatically obtain the most diverse statistical and substantive results from the transcribed and coded data, such as word frequency, lexical diversity and coherence, specific words and forms used in the respective speech session (such as: children's language errors as specific deviations from the norm of the respective language: units of

the so-called BABY TALK, onomatopoeia, overgeneralizations, children's and family occasionalisms, etc.).

CHILDES is an extremely flexible and convenient system, whose functionalities are available to any researcher even when the corpora are not yet published on the platform and when there is no access to the Internet. It is sufficient to have the programs of the CLAN package installed on the relevant computer and to have the necessary speech data transcribed in CHILDES terms. For example, long before the Bulgarian LabLing Corpus officially appeared, the system was used in this way in studies of the language development of Bulgarian children, which in turn was a kind of approbation of its applicability to our language. However, today, when the Bulgarian corpus is officially published, researchers have the freedom to choose whether to work online in the system itself, using the data from the Browsable transcripts section directly, or to download the data from the Download transcripts section to their personal device.

To illustrate this Table 1 presents the main steps in the algorithm of work directly in the platform.

Table 1 clearly shows that working directly in the platform is not only convenient, but also extremely easy. At the same time, the CHILDES system provides many opportunities, one of which is to examine all participants in the communication. This is clearly seen in Table 1 when comparing the sixth and seventh steps with the eighth and ninth steps, which extract data from the same transcript about the frequency of the modal verb must in the speech production of the examined child (SIM) and her mother (PLA). This is just one example of the broad applicability of the Bulgarian LabLing Corpus, which is predicated by the fact that each of the transcripts includes data on the identification of the researched persons (demographic and linguistic parameters) and on the type of the corresponding corpus (longitudinal or cross-sectional). See Fig. 3, Fig. 4:

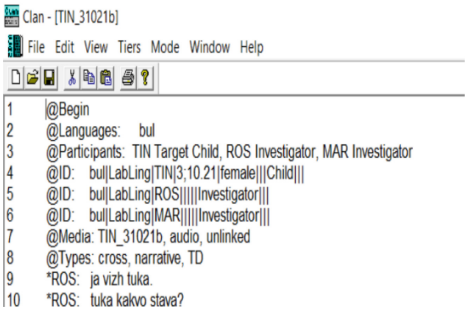
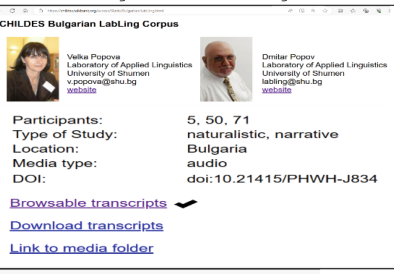
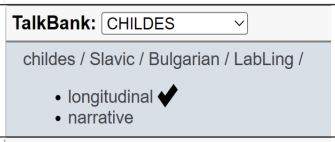
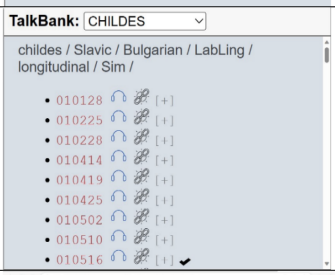
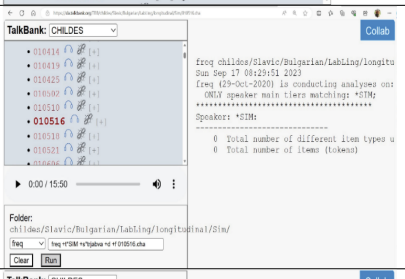
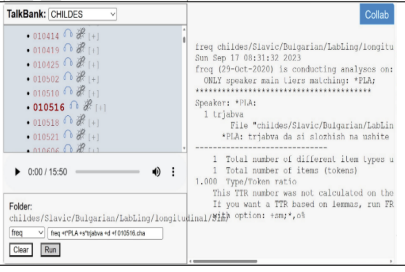
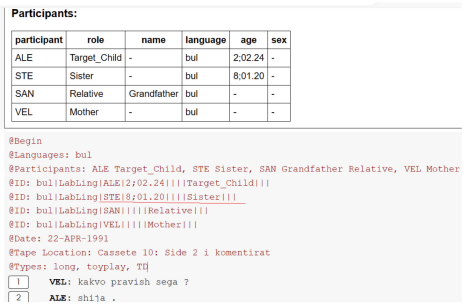


Fig. 3. Transcript lines in CHAT-format of cross-section data of Corpus B.

Table 1. Algorithm for working with CLAN programmes directly in CHILDES.

FIRST STEP **Access Bulgarian LabLing Corpus**	`https://childes.talkbank.org/access/Slavic/Bulgarian/LabLing.html`
SECOND STEP **Choose Browsable transcripts**	
THIRD STEP **Choose a corpus**	
FOURTH STEP **Choose a person**	
FIFTH STEP **Choose a transcript**	
SIXTH AND SEVENTH STEPS – **Enter a formula (CHILD):** `freq +t*SIM +s*trjabva +d +f 010516.cha` – **Result**	
EIGHTH AND NINTH STEPS – **Enter a formula (ADULT):** `freq +t*PLA +s*trjabva +d +f 010516.cha` – **Result**	



Participants:

participant	role	name	language	age	sex
ALE	Target_Child	-	bul	2;02.24	-
STE	Sister	-	bul	8;01.20	-
SAN	Relative	Grandfather	bul	-	-
VEL	Mother	-	bul	-	-

```
@Begin
@Languages: bul
@Participants: ALE Target_Child, STE Sister, SAN Grandfather Relative, VEL Mother
@ID: bul|LabLing|ALE|2;02.24||||Target_Child|||
@ID: bul|LabLing|STE|8;01.20||||Sister|||
@ID: bul|LabLing|SAN|||||Relative|||
@ID: bul|LabLing|VEL|||||Mother|||
@Date: 22-APR-1991
@Tape Location: Cassete 10: Side 2 i komentirat
@Types: long, toyplay, TU
1  VEL: kakvo pravish sega ?
2  ALE: shija .
```

Fig. 4. Identification data of researched persons in the transcript of the Bulgarian LabLing Corpus in the CHILDES System.

In this case, it is not simply a matter of simple identification of the researched persons, but of the possibility of extracting and analyzing the speech data of each one of them, or of a selected group, which is extremely important for ontogenetic studies. Access is also provided to the speech of adults to children, through which children receive information about the system of the target language (so-called positive input data). The results of such data analysis undoubtedly provide a solid basis for the construction of adequate explanatory models of language acquisition, in which it is possible to clearly distinguish the specific contribution of adults and their speech in the ontogenetic process in general, as well as in the acquisition of specific linguistic phenomena, and also to bring out the characteristics of the specific register called the speech of adults to children.

An additional research perspective can be found in study of the influence of the chosen child's siblings on the linguistic development of the respective studied child. In this regard, their metadata is of particular importance, which is also provided in the CHILDES system. Thus, among the participants of the interaction presented in Fig. 4 we find the elder sister (STE) of the examined child, for whom social role, gender, age, language are indicated.

In the logic of what has been said so far, it could be summarized that the very act of publishing Bulgarian data with child speech on the CHILDES platform is very important and useful for researchers, as it provides them with the opportunity to study the Bulgarian language in the context of an optimal working environment, characterized by free access to a wide variety of data, as well as computer programs through which the vast empirical array can be processed statistically.

At the same time, CHILDES offers one of the optimal electronic formats of modern corpus linguistics, enabling a multimodal integrated multi-aspect representation of speech behavior, which is a guarantee for obtaining objective results, as it provides the opportunity to observe the studied phenomenon simultaneously from different points of view. It is especially important in the study of the speaking child phenomenon due to the specific age characteristics and limited linguistic and communicative competence at a very early age.

In this regard, it is important to note the contributions of Brian MacWhinney [12,13] who developed the interactive multimodal system of the CHILDES platform specifically for the purposes of speech analysis and its optimal multi-aspect visualization, as a result of which a new quality is achieved in the studies of the components of non-verbal communication and the influence of the pragmatic context in child speech. The unified annotation of the extralinguistic data that accompanies the speech of the observed children, as well as the continuous mode of connection between the transcripts and the corresponding audio files, create opportunities and prospects not only for an isolated study of linguistic phenomena, but also for their in-depth comprehensive study in the dynamic mode of speech communication. With the help of CHILDES, the speech behavior of the studied children can be presented in a multimodal perspective.

McWhinney's innovative system for an integrated presentation of speech data is applicable to the Bulgarian corpus of child speech, since most of the transcripts are accompanied by audio files. At a later stage they could be synchronized, visualizing the executable audio file and its audible real-time sound implementation in SONIC MODE, displayed as an oscillogram at the bottom of the CLAN program window, as demonstrated in the sample in Fig. 5.

Fig. 5. Audiotranscript in SONIC MODE.

The option of multiple audio reconstruction of any fragment of the transcripts from the corpus guarantees a full and adequate analysis in the process of research of such complex linguistic phenomena and categories as, for example, modality, for the interpretation of which intonation is important, as well as of the phonetic side of child speech. In this regard, another advantage of CHILDES should be noted, namely the interoperability of its CLAN software package with other speech processing systems such as Praat, EXMARaLDA, and PHON.

5 Conclusion

The present article highlighted the benefits of the CHILDES interactive system. It is generally recognized that it provides conditions for greater precision in the collection,

transcribing and coding of data, and offers tools for automated analysis of large amounts of conversational material, which significantly expands the empirical base on which new theories are built. In addition, the unified convenient technique for annotating the extralinguistic data that accompanies the speech of the observed persons, and the continuous mode of connection between the transcripts and the corresponding audio and video files, create opportunities and perspectives for the study of all aspects of speech interaction.

With the publication of the Bulgarian data on the CHILDES platform the options for cross-linguistic research including another Slavic language are expanded. The Bulgarian linguistic tradition is enriched with another universal, convenient standard for the study of language ontogenesis, thanks to which comparisons with a large number of languages could be made quickly, accurately and reliably, on the basis of which relevant typologies and modern theories could be built.

The Bulgarian LabLing Corpus is only a separate miniature fragment of a multilingual virtual mosaic, which is constantly expanded and enriched within the framework of the two powerful systems CHILDES and TalkBank, which with their openness and rationality have established themselves as leaders in the processes of cooperation and globalization in the field of humanities research in general. This is a guarantee both for a broad social validity of research results based on their corpora, and for their integration into current work programs for the creation of infrastructures for the exchange of linguistic data and technologies aimed at overcoming the current fragmentation of research. In this sense, the integration of the American CMU-TalkBank into the European CLARIN (https://talkbank.org/knowledge/), a part of which is the first Bulgarian CHILDES corpus, created within the framework of ClaDa-BG, can be interpreted as a promising development.

Acknowledgments. The research presented in this paper is done within CLaDA-BG, Bulgarian national research infrastructure for resources and technologies for linguistic, cultural and historical heritage, integrated within CLARIN EU and DARIAH EU, funded by the Ministry of Education and Science of the Republic of Bulgaria. (support for the Bulgarian National Roadmap for Research Infrastructure).

References

1. Mladenov, St.: Nyakolko ezikoslovni vaprosi u prof. Iv. A. Georgov v rabotite mu za razvoya na detskiya govor. In God. na SU. IFF. kn. XXX. (1934). (In Bulgarian)
2. Popov, D.: Lingvistichna personologiya. Shumen, UI “Episkop Konstantin Preslavski” (2016). (In Bulgarian)
3. Popov, D.: Aspekti na balgarskoto proiznoshenie v sferite na standarta, substandarta i nonstandarta. In: Standart i substandart – diahronni i sinhronni aspekti. Shumen, UI “Episkop Konstantin Preslavski”, pp.191–204 (2005). (In Bulgarian)
4. Këster-Toma, S.: Standart, substandart, nonstandard. Rusistika **2**, 15–31 (1993)
5. Popov, D., Popova, V.: Fenomenat detska rech v svetlinata na mnogostepenniya model na balgarskata rech. – Otgovornostta pred ezika. Kniga 7. Shumen, UI “Episkop Konstantin Preslavski”, pp. 258–275 (2021). (in Bulgarian)
6. Stoyanova, Yu.: Problemi na psiholingvistikata. Sofya, UI “Sv. Kliment Ohridski” (2022). (in Bulgarian)

7. MacWhinney, B.: The CHILDES Project. Tools for Analyzing Talk, 2rd edn. Lawrence Erlbaum Ass., Mahwah, NJ (2000)
8. Hickmann, M.: Children's Discourse: Person, Space and Time Across Languages. Cambridge Studies in Linguistics, vol. 98. Cambridge University Press, Cambridge (2002)
9. Gülzow, I., Gagarina, N.: Intersentential pronominal reference in child and adult language. In: ZAS Papers in Linguistics, Nr. 48, pp. 203–223 (2007)
10. Gagarina, N., et al.: Multilingual Assessment Instrument for Narratives (MAIN). ZAS Papers in linguistics **56**, 1–140 (2012)
11. Gagarina, N. et al.: Multilingual Assessment Instrument for Narratives – Revised. ZAS Papers in Linguistics, 63 (2019)
12. MacWhinney, B.: Opening up video databases to collaborative commentary. In: Goldman, R., Pea, R., Barron, B., Derry, S. (eds.) Video research in the learning sciences, pp. 537–546. Lawrence Erlbaum Associates, Mahwah (2007)
13. MacWhinney, B., Wagner, J.: Transcribing, searching and data sharing: The CLAN software and the TalkBank data repository. In: Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion, Ausgabe 11, pp. 154–173 (2010)

Multiword Units in Russian Everyday Speech: Empirical Classification and Corpus-Based Studies

Natalia V. Bogdanova-Beglarian[1], Olga V. Blinova[1,2], Maria V. Khokhlova[1(✉)], Tatiana Y. Sherstinova[1,2], and Tatiana I. Popova[1]

[1] Saint Petersburg State University, Saint Petersburg, Russia
{n.bogdanova,o.blinova,m.khokhlova,t.sherstinova,t.i.popova}@spbu.ru
[2] HSE University, Saint Petersburg, Russia

Abstract. The article is dedicated to the results of a research project describing the classes and functioning of multiword units in contemporary Russian everyday speech. The concept of multiword units encompasses quite diverse linguistic phenomena, making the creation of a working typology one of the project's central tasks. This typology is necessary for annotating corpus material and obtaining statistical characteristics. The identified classes of multiword units include the following units: 1) non-phraseologized collocations, 2) phraseologized collocations, 3) occasional collocations, 4) idiom forms, 5) constructions, 6) precedent texts and their elements, 7) multi-word pragmatic markers, and 8) speech formulas. The article describes the methods for annotating these units using the ORD corpus of everyday spoken Russian and presents the results of a quantitative analysis of their functioning within the annotated subcorpus. The obtained data can be used to address both theoretical tasks in the field of lexical and grammatical description of Russian everyday speech and numerous tasks related to processing or generating live spoken Russian.

Keywords: Modern Russian · Everyday Speech · Oral Discourse · Multiword Units · Collocations · Syntax · Statistical Analysis · Speech Corpus · Corpus Linguistics · Speech Technologies

1 Introduction

The study of spoken language, especially using a corpus approach based on recordings obtained in natural communication settings, reveals phenomena that are not reflected in existing dictionaries and grammars but actively function in the speech of native speakers. These phenomena, therefore, require special documentation and analysis for linguistic and language-teaching purposes, as well as for creating human-like dialogue systems and artificial intelligence. This research focuses on phenomena at the intersection of vocabulary, grammar, and syntax, which we refer to collectively as *multiword units*. A number of works are dedicated to their analysis and study, for example, [1–10]. These

A. Karpov and V. Delić (Eds.): SPECOM 2024, LNAI 15299, pp. 187–200, 2025.
https://doi.org/10.1007/978-3-031-77961-9_14

units require not only theoretical description but also the formation of an inventory—a lexicon (in the broadest sense of the word), in which they would be represented with the necessary quantitative characteristics. For the codified Russian written language, multiword units are relatively fully described. However, in the class of multiword units used in spontaneous spoken speech, there are still many "white spots" despite the emerging linguistic and digital resources. For example, databases exist for the Russian language that combine units of different types: multiword units, collocations, constructions, etc. ([11–17], etc.).

The relevance of addressing the problem outlined is connected to the fact that, recently, linguistics has become closely intertwined with information technology and the development of various speech applications. Solving these tasks requires not only a coherent theory but also a large volume of annotated linguistic data. The study of multiword units and their identification in a speech corpus involves addressing issues related to lemmatization and the representation of their morphological, syntactic, and semantic features.

The material for studying multiword units is the ORD corpus of Russian everyday speech, characterized by the fact that recordings are obtained in natural communicative situations [18, 19] and reflect the full richness and diversity of everyday speech communication—in terms of the topics of conversation [20], participants [21], and communication conditions [22].

The approaches to searching for and identifying multiword units are diverse, relying on both expert and automatic techniques. In our study, it seems reasonable to employ both. The initial list of multiword units was obtained through expert methods, followed by n-gram analysis of transcriptions of everyday spoken language to identify the most frequent bigrams and trigrams. The results of the n-gram analysis were described in [23, 24]. However, the overall number of n-grams obtained, amounting to tens of thousands of units, and the lack of context pose an obvious drawback for subsequent expert work. Therefore, in this study, the basis for collecting multiword units was their expert manual annotation, described below in Sect. 3, and relying on their empirical classification presented in Sect. 2. The article also presents the results of automatic clustering of the empirically obtained list of units (see Sect. 4). Finally, Sect. 5 provides preliminary statistics on the distribution of multiword units based on the study sample.

2 Empiric Classification of Multiword Units

The concept of *multiword units* encompasses a wide range of linguistic phenomena, so the creation of their working typology is one of the central tasks of the project. This typology is essential for the subsequent annotation and processing of the material. The typology of multiword units was developed in several iterations. In the first stage, a pilot classification of multiword units was used, focusing on their structural and lexical features. Based on the results of comprehensive pilot annotation of oral speech transcriptions, taking into account the main proposed types of multiword units, this typology was revised, and a new scheme was proposed. This scheme currently includes eight main categories:

1. Non-phraseologized collocations,
2. Phraseologized collocations,

3. Occasional collocations,
4. Idiom forms,
5. Constructions,
6. Precedent texts and their elements,
7. Multi-word pragmatic markers,
8. Speech formulas.

Non-phraseologized collocations are stable combinations whose perception does not determine the imagery of the meaning.

Phraseologized collocations are stable constructs whose elements possess figurative meanings. As a result of the interaction between the semantics of the construction components, a certain meaning is fixed in spoken language for the unit: "ne obrashchat' vnimaniya" ("to ignore"), "v poryadke veshchey" ("as a matter of course"), "doyti do ruchki" ("to reach the limit"). This type of multiword unit is closest to traditional phraseological units.

Occasional collocations, as the name suggests, are modifications of commonly accepted collocations in the language.

An **idiom form** is considered a word form that, due to frequent use, acquires functional and semantic significance in everyday communication (most often this is a prepositional-case form of nouns): for example, "po ponyatiyam" ("according to the rules"), "do figa" ("a lot"), "ne v kaif" ("not enjoyable"), "v printsipe" ("in principle"), "ne po sebe" ("uncomfortable"), "ne gorit" ("not urgent"), and others.

The concept of a **construction** differs from an idiom form and a phraseologized collocation in that the structure of the construction includes a constant component and a variable component X: < X ni razu ne Y > ("X never Y"), < X-u ne do Y-a > ("X doesn't care about Y"), < nu + Acc! > ("come on + Acc!"), etc.

Elements of precedent texts refer to fragments of well-known phrases, for example, from movies: "ikh yest' u menya" ("I have them") (a phrase from Lev Slavin's play "Intervention"), "chey tuflya" ("whose shoe") (from Leonid Gayday film "Kidnapping, Caucasian Style"), etc.

Pragmatic markers are functional units of oral discourse that help speakers structure dialogue and mark speech intention. Pragmatic markers often have a complex structure, thereby expanding the list of multiword units: "ya ne znayu" ("I don't know"), "skazhem tak" ("let's say"), "kak govorit'sya" ("as they say"), "tak skazat'" ("so to speak"), "nu vot" ("well then"), "i vse dela" ("and all that"), "ili kak eto" ("or whatever"), etc.

Speech formulas are often interjectional units that reflect the speaker's emotional reaction or a response in a dialogue: "vot yeshchyo!" ("there you go! "), "nichego sebe!" ("wow!"), "kak khochesh" ("as you wish"), "kak znayesh" ("as you know").

The results of the pilot annotation showed that the proposed classification generally well reflects the features of multiword units characteristic of spoken language, therefore it is accepted as the main one for conducting expert annotation of these units and further research.

3 Expert Annotation of Multiword Units in ORD Corpus

The ORD corpus is a complex and multi-component resource used for conducting research on Russian spoken discourse at all linguistic levels. An important result of the ongoing research is the manual expert annotation of the corpus materials at the level of multiword units.

3.1 Multiword Units Annotation Principles

The annotation of multiword units in the ORD corpus is carried out as follows. Experts review the transcriptions of audio recordings, which are exported into a tabular format, and fill in the multiword units database using a form that includes the following fields:

1. Communicative episode,
2. Speaker code,
3. Phrase,
4. Multiword unit as it appears in the text,
5. Class of the multiword unit according to the proposed typology (the *Tags* column in the database),
6. Invariant (optional—filled in only when the form of the initial multiword unit was clear without doubt),
7. New multiword unit—a note indicating whether the multiword unit was included in the initial list.

The following annotation codes were proposed:

1. Non-phraseologized collocations—NK,
2. Phraseologized collocations—FK,
3. Occasional (non-conventional) collocations—OK,
4. Idiom forms—ID,
5. Constructions—KS,
6. Elements of precedent texts—PT,
7. Multi-word pragmatic markers—PM,
8. Speech formulas—RF.

In the *Tags* column, it was preferable to record only one (the main) variant of the multiword unit's characteristic, as, unlike pragmatic markers of spoken language, the units under study are not prone to multifunctionality; they are primarily annotated in terms of their formal organization.

3.2 Multiword Units Annotation Results

Four experts participated in the annotation process, one of whom (E1) acted as the curator and made final corrections. The episodes of natural speech communication selected for annotation varied in duration, and thus, differed in the volume of text transcriptions. Therefore, multiple speech episodes were selected for some informants, affecting the distribution of material among the experts for annotation. Each episode was annotated by one expert. The final distribution of annotated episodes was as follows:

– E1 – 14 episodes (7.8%);
– E2 – 20 episodes (10.26%);
– E3 – 101 episodes (51.79%);
– E4 – 60 episodes (30.77%).

The experts thoroughly reviewed the text transcriptions in the *Phrase* column of the research database and recorded information about the multiword units found in the phrases in columns created specifically for annotation. The multiword units were recorded in the form they appeared in the fragment in a specially designated column (*Multiword units*).

In total, 195 macro-episodes were annotated, with a total volume of 300,000 word usages. The manual annotation of multiword units enabled the creation of an expanded list of these units, preliminary statistical information on the implementation of these units in spoken language, and the identification of the main difficulties in expert annotation of these units.

3.3 Main Challenges in the Annotation Process of Multiword Units

The main difficulties in the annotation process arose with those multiword units that were not included in the initial list, necessitating collective decisions on the classification of each unit and whether the unit could be considered a multiword unit at all.

A particularly close connection was found between constructions and non-phraseologized collocations, as the lack of imagery and figurative meaning of the components distanced the unit under study from phraseologized collocations. In such cases, the determining factor was the search for a variable (X) that is defining from the conceptual framework's perspective.

Another problem in the annotation of multiword units was that some word combinations primarily realized their grammatical meaning and syntactic valence rather than stability and lexicalization, which prevented them from being classified as multiword units, even though their combination could potentially be considered regular.

At the final stage of annotation, it became clear that identifying the invariant for each realization of multiword units is a separate research task. For example, in constructions, it was necessary to determine the fixed part and its form, and then the part that is variable. Next, the grammatical and lexical characteristics of the potential variable needed to be described.

In speech formulas, some components may be perceived as optional, but upon analysis, it becomes clear that only the full composition of the multiword unit realizes its meaning. For example, the unit "da ladno" ("oh, come on") can serve as a reaction precisely in this structural variant because identifying "da" ("oh") as an optional part "(da) ladno" ("come on") causes the unit to cease being a multiword unit and loses its function of expressing the speaker's reaction.

Some units annotated as phraseologized collocations also exhibit variability. For example, the multiword units "morochit' golovu" ("to mess with someone's head") and "vynosyt' mozgi" ("to blow someone's mind") can be perceived as synonymous, or their proximity can be seen as a potential to fill positions with words of a certain meaning, allowing the multiword units to be considered as constructions. Only further expert

work and linguistic analysis will allow the formation of a final list of invariants for each realization of multiword units and the creation of a new classification of multiword units in terms of their formal organization.

4 Automatic Clustering of Multiword Units

The empirically derived list of multiword units was subsequently subjected to an automatic clustering procedure. Initially, automatic clustering of multiword units was carried out based on the results of expert annotation of oral speech transcriptions for a sample of 300,000 tokens. The clustering was performed using the k-means algorithm without considering metadata but utilizing two different approaches to data vectorization: 1) tf-idf (CountVectorizer from sklearn) and 2) FastText embeddings. Calculations were performed for models with 5, 10, 15, and 30 clusters[1]. During automatic clustering, the elements within each cluster were grouped around one or more keyword features.

The most semantically meaningful clusters were obtained when the sample was divided into 30 clusters. For example:

CLUSTER #8 multiword units:

[‘*million raz’ (‘a million times’), ‘desyat’ raz’ (‘ten times’), ‘inoy raz’ (‘sometimes’), ‘pervyy raz slyshu’ (‘first time I hear it’), ‘sto raz’ (‘a hundred times’), ‘paru raz’ (‘a couple of times’), ‘kak raz’ (‘just right’), ‘lishniy raz’ (‘one more time’)*].

CLUSTER #22 multiword units:

[‘*ne moyo’ (‘not my thing’), ‘ne pozhalela deneg’ (‘didn’t spare the money’), ‘ne problema’ (‘no problem’), ‘ryadom ne stoyat’ (‘don’t come close’), ‘darom ne nuzhna’ (‘don’t need it for free’), ‘ne bum-bum’ (‘don’t get it’), ‘ne svetit’ (‘not gonna happen’), ‘ne sud’ba’ (‘not meant to be’), ‘ne govorite’ (‘don’t say’), ‘sovest’ ne gryizla’ (‘didn’t feel guilty’), ‘nikak ne doberus” (‘can’t get around to it’)*].

Nevertheless, automatic clusters can sometimes contain an “exception”. For example, cluster #19 (when dividing multiword units into 30 clusters) mainly consists of units containing the lemma “delo” (“thing” or “matter”). However, for some reason, the borrowed English multiword unit “vi a ze chempions” (“we are the champions”), which belongs to the type of precedent texts, also ended up in this cluster.

CLUSTER #19 multiword units:

[‘*odno delo’ (‘one thing’), ‘sovsem drugoe delo’ (‘a completely different matter’), ‘obychnoe delo’ (‘a usual thing’), ‘imeyu delo’ (‘have a matter’), ‘khoroshee delo’ (‘a good thing’), ‘delo khoroshee’ (‘the matter is good’), ‘takie dela’ (‘such things’), ‘ponyatnoe delo’ (‘obviously’), ‘poslednee delo’ (‘the last thing’), ‘delo poshlo’ (‘the matter progressed’), ‘strannoe delo’ (‘a strange thing’), ‘takoe delo’ (‘such a thing’), ‘sereznoe delo’ (‘a serious matter’), ‘svyatoe delo’ (‘a sacred thing’), ‘vi a ze chempions’ (‘we are the champions’), ‘temnye dela’ (‘dark matters’), ‘drugoe delo’ (‘another matter’), ‘bylo delo’ (‘there was a matter’)*].

[1] The choice of the maximum value of the number of clusters depends on the volume of the analysed data. In the first experiment the results of manual annotation of multiword units were clustered, while in the second experiment the lists of n-grams obtained automatically were processed, hence, the volume of data in the second case was larger.

At the next stage of data processing, automatic clustering was performed for the complete list of frequent n-grams, where *n* takes a value from 2 to 5 for the entire volume of existing oral speech transcripts of the ORD corpus. The clustering was conducted using the k-means algorithm without considering metadata, but utilizing two different approaches to data vectorization: tf-idf and FastText embeddings. Given the large number of units studied, amounting to tens of thousands of unique types for each of the 2-, 3-, 4-, and 5-grams, it was decided to divide the research sample into 50 clusters.

For each n-gram size, four files were obtained—two txt files with clusters (where key features are highlighted as a separate line for the tf-idf model) and two csv files with complete lists of types and the cluster number in the second column. These tables allow for the analysis of statistics and, if necessary, enable the data to be traced back to the original sources.

The results show that the n-gram clusters differ significantly from each other in structural and semantic cohesion. See, for example, clusters 1 and 45 for bigrams:

CLUSTER #1.

Types: [*'ya priedu' ('I will come'), 'ya zabyl' ('I forgot'), 'ya rabotayu' ('I am working'), 'ya poprobuyu' ('I will try'), 'ya chitayu' ('I am reading'), 'ya reshil' ('I decided'), 'ya yezdila' ('I went'), 'ya vspomnila' ('I remembered'), 'ya rad' ('I am glad'), 'ya kupila' ('I bought'), 'ya skhozhu' ('I will go'), 'ya ya' ('I I'), 'ya poprosil' ('I asked'), 'kak ya' ('like I'), 'ya vozmu' ('I will take'), 'ya yeye' ('I her'), 'ya polozhila' ('I put'), 'naskol'ko ya' ('as far as I'), 'ya napishu' ('I will write'), 'kotoroye ya' ('which I')*].

CLUSTER #45.

Types: [*'ugu kogda' ('mm-hmm when'), 'mam a' ('mom uh'), 'nado tuda' ('need to go there'), 'on yey' ('he her'), 'zdes' bylo' ('here was'), 'ya dolzhen' ('I must'), 'podozhdi a' ('wait uh'), 'vidite kak' ('see how'), 'togda davay' ('then let's'), 'tri shtuki' ('three pieces'), 'interesno ya' ('interesting I'), 'dumala ya' ('thought I'), 'moemu a' ('my uh'), 'ta m' ('that um'), 'a yey' ('uh her'), 'ya im' ('I them'), 'ugu u' ('mm-hmm uh'), 'e potomu' ('uh because'), 'ponimayesh' ty' ('you see'), 'vy tozhe' ('you too')*].

Even greater diversity is observed for larger n-grams. The conclusion that can be drawn from this study is that, in the future, clustering should be performed not on the entire array of n-grams obtained, but only on the most frequent units (the upper zone of the n-gram frequency dictionary).

The study showed that automatic clustering, with a correctly selected number of classes, is a useful tool for the preliminary grouping of word sequences based on their lexicon. Since automatic clustering relies solely on the lexical composition of multiword units without considering semantics, expert analysis is necessary for further work with such data. A useful property of a cluster is its reliance on "key" word(s), which allows grouping similar multiword units and can be used to search for invariant forms. For example:

CLUSTER #30 (when dividing multiword units into 30 clusters).

multiword units: [*'vot eto vot' ('this one here'), 'vot ona vot' ('here she is'), 'vot eti vot' ('these ones here'), 'vot beda' ('what a trouble'), 'vot etot vot' ('this one here'), 'vot tebe' ('here you go'), 'vot eti samye' ('these very ones'), 'vot eta bol' vot' ('this pain*

here'), 'vot takiye vot dela' ('that's how things are'), 'vot takoy vot' ('this kind of'), 'vot etu vot' ('this one here'), 'vot takiye vot' ('these kinds of'), 'vot takiye dela' ('these are the things'), 'vot tuda vot' ('over there'), 'vot takaya vot' ('this kind of'), 'vot imenno' ('exactly'), 'vot etim vot' ('with these here'), 'vot tak vot' ('that's how it is')].

It can be assumed that this property of clusters will be most pronounced with a sufficiently large number of them. However, this hypothesis requires experimental verification.

Regarding the clustering of a large number of automatically obtained n-grams, the analysis showed that the results do not have a distinguishing function that would be useful for the automatic identification of multiword units, at least for the counting methodology used in the project. This problem might be resolved by machine learning methods based on expert selection, but for this task, the volume of expert annotation needs to be significantly expanded.

5 Preliminary Statistics of Multiword Units Distribution in Everyday Conversations

In the course of the study, statistical data were obtained on the conditions of the realization of multiword units in everyday spoken language and their distribution in specific types of communicative macro-episodes, as well as in relation to other communication conditions. A description of the obtained statistics was also provided.

5.1 Most Frequent Multiword Units

The overall frequency of use of multiword units in a representative sample was obtained. The total lexicon of multiword units identified from the ORD material during manual annotation (on a subsample of 300,000 words, 195 speech episodes) amounted to 1,088 units of various types (see Sect. 2).

The results showed that the composition of these most frequent stable multiword units in our everyday communication is quite heterogeneous.

The most frequent unit "V PRINTSIPE" ("in principle") (rank 1) is a lexicalized prepositional-case form (idiom form) or a pragmatic marker (verbal hesitative or delimiter, primarily navigational, depending on the context).

Similarly, the unit "V OBSHCHEM" ("generally") (rank 10) in this frequency list can be characterized. The idiom form "V OBSHCHEM" ("generally") as a pragmatic marker is a verbal hesitative, delimiter of all three types (initial, navigational, and final), and occasionally a self-correction marker, also depending on the context.

From the class of pragmatic markers in the top 10, there are also units "ETO SAMOE" ("you know") (rank 2) (verbal hesitative, self-correction marker, delimiter marker of all three types (initial, navigational, and final), and rarely a xenopointer marker), "NA SAMOM DELE" ("actually") (rank 5) (verbal hesitative), and "I TAK DALEE" ("and so on") (rank 7) (placeholder marker).

Thus, 50% (exactly half) of the most frequent multiword units in our spoken communication are primarily pragmatic markers, which are not included in this status in traditional explanatory dictionaries, including dictionaries of Russian colloquial speech,

nor in the "Russkiy konstruktikon" [25], nor in the "Pragmatikon" [26]. All data on these markers (their functional characteristics) are provided here according to the Dictionary of Pragmatic Markers [17]. Two of the 5 units of this type ("V PRINTSIPE" ("in principle") and "V OBSHCHEM" ("generally")) are also idiom forms, constituting a separate class of multiword units. This polyfunctionality is characteristic of many spoken language units, reflecting the overall diffuse nature of this material.

The remaining units that made it into the top 10 are speech formulas (40%) ("NICHEGO SEBE" ("wow"), "SLAVA BOGU" ("thank God"), "DA TY CHO" ("really"), "VSE RAVNO" ("anyway")), included in the "Pragmatikon" since they are predominantly response replicas in dialogue, and a phraseologized collocation (10%) ("VSE VREMYA" ("all the time")), definitely included in the "Russkiy konstruktikon".

It should also be noted that all the most frequent multiword units in our everyday speech are bi- and trigrams, described in [23]; [24].

5.2 Most Frequent Classes of Multiword Units

The top 5 of this frequency list include phraseologized collocations (rank 1), idiom forms (rank 2), speech formulas (rank 3), pragmatic markers (rank 4), and syntactic constructions (rank 5) ("delo v tom chto" ("the fact is that"), "v lyubom sluchaye" ("in any case"), etc.).

The analysis showed that among the phraseologized collocations, the most commonly used units in everyday Russian speech are "VSE VREMYA" ("all the time") (3.67%) and "PONYATNOE DELO" ("obviously") (2.20%) (percentage calculated within each group); the most frequent idiom form is "V PRINTSIPE" ("in principle") (34.55%).

For speech formulas, the top 4 ranks include the same units "NICHEGO SEBE" ("wow"), "SLAVA BOGU" ("thank God"), "DA TY CHO" ("really"), and "VSE RAVNO" ("anyway"), which are in the top 10 of the overall frequency list of multiword units.

For the group of pragmatic markers (PM), again, the top 4 positions are occupied by units from the overall frequency list of multiword units ("ETO SAMOE" ("you know"), "NA SAMOM DELE" ("actually"), "I TAK DALEE" ("and so on"), "V OBSHCHEM" ("generally")). It is also evident that the obtained data reflect the frequency of realizations of multiword units, not their base variants (invariants). In the lexicon of Russian Pragmatic Markers [17], the realizations "ETO SAMOE" and "ETOT SAMYI" are one marker "ETO SAMOE" ("you know") (this "classic" form is the most frequent in our speech and is used in any hesitative search, including when grammatical adjustment to the desired noun is not required); the realizations "VOT TAK VOT" and "VOT ETO VOT" are also one deictic marker "VOT (…) VOT" ("this one here"), which exists exclusively as a structural model that is filled each time with a new unit: "VOT TAK VOT" ("this way"), "VOT TAKOY VOT" ("this kind of"), "VOT OTSYUDA VOT" ("from here"), etc. This marker simply does not have a single base (standard) form, which is why it occupies a special place in the lexicon of pragmatic markers. Neither dictionaries nor grammars of the Russian language highlight this construction as an independent unit, whereas corpus material analysis shows its very high frequency (rank 19 in the list of 60 Russian pragmatic markers).

Among syntactic multiword unit constructions, the most common are “V LYUBOM SLUCHAYE” (“in any case”) and “DELO V TOM CHTO” (“the fact is that”) (4.88% each), among non-phraseologized collocations are “ODNU SEKUNDOCHKU” (“one moment”) (5.38%), as well as “DRUGOE DELO” (“another matter”), “PO KRAYNEY MERE” (“at least”), and “CHEGO-TO TAKOE” (“something like that”) (4.62% each). Again, it is clear that this refers only to specific realizations of multiword units. For example, alongside “ODNU SEKUNDOCHKU” the lexicon contains “ODNU SEKUNDU” (“one second”) (rank 56). However, the expected invariant form “CHTO-TO TAKOE” (“something like that”) was not found next to “CHEGO-TO TAKOE”. This once again indicates that the question of multiword unit invariants is not as simple as it seems at first glance and requires separate consideration.

Multiword units from the classes of occasional collocations and precedent texts are predictably rare. Interestingly, a significant portion of occasional collocations units include obscene vocabulary, although such vocabulary is also present in other groups. Overall, both of these classes of multiword units provide good material for analysis from various perspectives.

5.3 Part-Of-Speech Composition of Multiword Units

The entire material of the annotated subcorpus (a subsample of 300,000 words from 195 speech episodes) was automatically tagged for the part-of-speech (POS) of the components of multiword units, allowing for the generation of frequency lists based on this parameter.

The most frequent POS structure turned out to be PREP NOUN (a noun with a preposition, lexicalized prepositional-case word form, or idiom form) (14.85%). The most typical units of this type are: “V PRINTSIPE” (“in principle”) (40.59%), “V SMYSLE” (“I mean”) (5.61%), “V ITOGE” (“as a result”) (3.96%), “PO IDEE” (“supposedly”) (3.63%).

Other frequent structures are ADJF NOUN (a combination of a full adjective (including adjective-pronoun and numeral-pronoun) with a noun) (5.98%) and PREP ADJF NOUN (the same combination with a preposition) (5.78%). The most typical units of these two types are: “PONYATNOE DELO” (“obviously”) (9.84%), “ODNU SEKUNDOCHKU” (“one moment”) (5.74%), “DRUGOE DELO” (“another matter”) and “KAKAYA RAZNITSA” (“what’s the difference”) (4.92% each); “NA SAMOM DELE” (“actually”) (23.73%), “V LYUBOM SLUCHAE” (“in any case”) (11.86%), “DO SIKH POR” (“up to now”) (6.78%), “VO VSYAKOM SLUCHAE” (“anyway”) and “PO KRAYNEY MERE” (“at least”) (5.08% each).

5.4 Frequency of Use of Multiword Units Depending on Speakers’ Social Characteristics

The research sample included speech episodes from 111 informants’ speech days, among which there were 57 women and 64 men. The sample also included the speech of their 727 interlocutors, among which there were 645 women and 272 men. More than 50% of the material studied involved domestic communication, with business communication being the second most frequent.

These data correlate with information about the social roles of the speakers: most often, speakers took on the role of "friend", with the second most common role being "work colleague".

It has already been noted that multiword units from occasional collocations and precedent texts classes are predictably rare. Interestingly, occasional collocations multiword units (33 instances in the material) are used equally by both women and men: 17 uses in women's speech and 16 in men's speech. Of the 23 instances of precedent texts, 14 are used by women and only 9 by men.

The use of multiword units from the Non-phraseologized collocations, constructions, and pragmatic markers classes is relatively evenly distributed among women (60% of uses) and men (40% of uses). The use of phraseologized collocations is slightly more common among women (55%), while speech formulas are more characteristic of women's speech (68%).

The distribution of multiword units across age groups does not have striking features, as the percentage distribution is relatively even. Only a few indicators stand out:

– Older men use idiom forms less than middle and younger age groups;
– In the speech of older women, speech formulas are predominant.

The level of speech competence is determined in the ORD corpus through the correlation of two indicators: the level of education and the professional activity of the informant. The results indicate that the use of various classes of multiword units is generally more characteristic of people with an intermediate level of speech competence (only 5 to 20% among people with a high level of speech competence).

Other features of the use of multiword units in different communication situations were also identified and described.

6 Conclusion

The study presents a typology of multiword units for spontaneous everyday Russian speech and provides statistical data on their realization based on the manually annotated subcorpus of the well-known ORD corpus. Due to the labor-intensive nature of manual annotation, only one-third of the existing transcriptions in the corpus have been annotated to date. Therefore, the presented statistics should be considered preliminary, and the study of multiword units continues along the following paths: 1) by expanding the volume of annotated data to 1 million word usages and 2) by involving automatic analysis tools for processing multiword units [27].

Methods for automatically identifying multiword units will rely on existing lexicons, but due to the homonymy of linguistic units, they will require subsequent manual correction. Special scripts are being created to search for new forms of constructions [28] based on invariant structures of multiword units. In addition, modern speech technologies allow for a significant expansion of the empirical base of corpus research by attracting new representative volumes of audio recordings. Such work is currently being carried out on the materials of the ORD corpus [29]; [30], and conducting statistical analysis of multiword units on extended volumes of transcriptions will allow for the correction of quantitative data on their usage in different communication situations by different types of speakers.

The obtained data can be used to address both theoretical tasks in the field of lexical and grammatical description of Russian everyday speech and numerous tasks related to processing or generating live spoken Russian. Additionally, the research results will form the basis of a Dictionary of Collocations and other multiword units of everyday Russian speech.

Acknowledgments. This research has been carried out thanks to the financial support of Russian Science Foundation (project No. 22–18-00189 "Structure and Functionality of Stable Multiword Units in Russian Everyday Speech").

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Columbus, G.: Processing MWUs: Are different types of MWUs psycholinguistically-valid? An eye-tracking study. In: Wood, D. (ed.) Perspectives on formulaic language incommunication and acquisition, pp. 194–210. Continuum, New York (2010)
2. Moon, R.: Vocabulary Connections: Multi-word Items in English. In: Schmitt, N., McCarthy, M. (eds.) Vocabulary: Description, Acquisition and Pedagogy, pp. 40–63. Cambridge University Press, Cambridge (1997)
3. Moon, R.: Frequencies and forms of phrasal lexemes in English. In: Cowie, A.P. (ed.), Phraseology: Theory, analysis, and applications, pp. 79–100. Clarendon Press, Oxford (1998)
4. Nattinger, J., DeCarrico, J.: Lexical phrases and language teaching. Oxford University Press, Oxford (1992)
5. Nunberg, G., Sag, I., Wasow, Th.: Idioms. Language **70**(3), 491–538 (1994)
6. Schweigert, W.: The comprehension of familiar and less familiar idioms. J. Psycholinguist. Res. **15**, 33–45 (1986)
7. Weinreich, U.: Problems in the analysis of idioms. In: Puhvel, J. (ed.), Substance and structure of language, pp. 23–81. University of California Press, Berkley (1969)
8. Wray, A.: Formulaic language and the lexicon. Cambridge University Press, Cambridge (2002)
9. Wray, A.: Formulaic language: Pushing the boundaries. Oxford University Press, Oxford (2008)
10. Bogdanova-Beglarian, N., Blinova, O., Khokhlova, M., Sherstinova, T.: Towards the description of multiword units in Russian everyday speech: state-of-the-art and the methodology of further research. In: Bolgov, R., Mukhamediev, R., Pereira, R., Mityagin, S. (eds.) Digital Geography. IMS 2022, pp. 129–139. Springer Geography. Springer, Cham (2024)
11. Bast, R., et al: The Russian Constructicon. An electronic database of the Russian grammatical constructions. (2021) https://constructicon.github.io/russian/. Accessed 15 July 2024
12. Janda, L.A., Lyashevskaya, O., Nesset, T., Rakhilina, E., Tyers, F. M.: Chapter 6. A constructicon for Russian: filling in the gaps. In: Lyngfelt, B. et al. (eds.), Constructicography: Constructicon development across languages [Constructional Approaches to Language 22], pp. 165–181. John Benjamins Publishing Co., Amsterdam (2018). https://doi.org/10.1075/cal.22.06jan

13. Khokhlova, M.: Collocations in Russian lexicography and Russian collocations database. In: Proceedings of the 12th Language Resources and Evaluation Conference, pp. 3198–3206. ELRA, Marseille, France (2020)
14. Kopotev, M., Escoter, L., Kormacheva, D., Pierce, M., Pivovarova, L., Yangarber, R.: CoCoCo: online extraction of Russian multiword expressions. In: The 5th Workshop on Balto-Slavic Natural Language Processing (10–11 September 2015, Hissar, Bulgaria), pp. 43–45. INCOMA Ltd, Sofia (2015)
15. Khokhlova, M.: Attributive collocations in the gold standard of Russian collocability and their representation in dictionaries and corpora. Voprosy Leksikografii **21**, 33–68 (2021)
16. Lyashevskaya, O., Kashkin, E.: FrameBank: a database of Russian lexical constructions. In: M.Yu. Khachay, N. Konstantinova, A. Panchenko, D.I. Ignatov, G.V. Labunets (eds.), Analysis of Images, Social Networks and Texts. Fourth International Conference, AIST 2015, Yekaterinburg, Russia, April 9–11, 2015, Revised Selected Papers. Communications in Computer and Information Science, Vol. 542, Springer, pp. 337–348 (2015)
17. Pragmatic markers of Russian everyday speech: Dictionary-monograph / Ed. N.V. Bogdanova-Beglaryan. St. Petersburg: Nestor-History (2021)
18. Asinovsky, A., Bogdanova, N., Rusakova, M., Stepanova, S., Ryko, A., Sherstinova, T.: The ORD speech corpus of Russian everyday communication "one speaker's day": creation principles and annotation, Lecture Notes in Computer Science – Vol. Text, Speech and Dialogue, no. 5729/2009. (2009)
19. Sherstinova, T.: The structure of the ORD speech corpus of Russian everyday communication. In: Matoušek, V., Mautner, P. (eds.) TSD 2009. LNAI, vol. 5729, pp. 258–265. Springer, Berlin-Heidelberg (2009)
20. Akinshina, E., Sherstinova, T.: Thematic diversity of everyday Russian Discourse: a case study based on the ORD corpus. In: Mahadeva Prasanna et al. (eds.), Specom 2022, LNCS 13721, pp. 1–9. Springer, Cham (2022)
21. Bogdanova-Beglarian, N., et al.: Sociolinguistic extension of the ORD corpus of Russian everyday speech. In: Ronzhin, A. et al. (eds.) SPECOM 2016, Lecture Notes in Artificial Intelligence, LNAI, vol. 9811, pp. 659–666. Springer, Switzerland (2016)
22. Sherstinova, T.: Macro episodes of Russian everyday oral communication: towards pragmatic annotation of the ORD speech corpus. In: Ronzhin, A. et al. (eds.) SPECOM 2015, Lecture Notes in Artificial Intelligence, LNAI, vol. 9319, pp. 268–276 (2015)
23. Khokhlova, M., Blinova, O., Bogdanova-Beglarian, N., Sherstinova, T.: On the most frequent sequences of words in Russian spoken everyday language (bigrams and trigrams): an experience of classification. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). SPECOM 2023, 14338 LNAI, pp. 455–466 (2023)
24. Sherstinova, T., Markovich, O.: N-gram analysis of everyday Russian speech: in search of multiword units. In: 35th Conference of Open Innovations Association (FRUCT), pp. 831–838. Tampere, Finland (2024)
25. Russkiy konstruktikon: https://constructicon.github.io/russian/
26. Pragmatikon: https://pragmaticon.ruscorpora.ru
27. Sherstinova, T., Popova, T.: Multiword units in Russian spontaneous spoken language: methods for lexicon expansion and statistical analysis. LiLaC-2024 (2024) (in print)
28. Rakhilina, E. V. (ed.): Lingvistika konstrukciy. Azbukovnik Publishing Center, Moscow (2010)

29. Sherstinova, T., Kolobov, R., Mikhaylovskiy, N.: Everyday conversations: a comparative study of expert transcriptions and ASR outputs at a lexical level. In: Proceedings of SPECOM 2023, LNCS, 14338/14339, pp. 43–56 (2023)
30. Sherstinova, T., Mikhaylovskiy, N., Kolpashchikova, E., Kruglikova, V.: Bridging gaps in Russian language processing: AI and everyday conversations. In: 35th Conference of Open Innovations Association (FRUCT), pp. 253–258. Tampere, Finland (2024)

Neurophysiological Correlates of Textual Modulation in Visual Stimuli: An Experimental Study of Russian and English Memes

Rodmonga Potapova[1], Vsevolod Potapov[2], Ekaterina Karimova[3], Leonid Motovskikh[1], and Nikolay Bobrov[1(✉)]

[1] Institute of Applied and Mathematical Linguistics, Moscow State Linguistic University, 38 Ostozhenka Street, 119034 Moscow, Russia
RKPotapova@yandex.ru, leon@motovskikh.ru, Arctangent@yandex.ru

[2] Centre of New Technologies for Humanities, Lomonosov Moscow State University, Leninskije Gory 1, 119991 Moscow, Russia
volikpotapov@gmail.com

[3] Laboratory of Applied Physiology of Human Higher Nervous Activity, Institute of Higher Nervous Activity and Neurophysiology of RAS, 5A Butlerova street, 117485 Moscow, Russia
e.d.karimova@gmail.com

Abstract. In this pilot work neurophysiological correlates of textual modulation of perception of visual stimuli using English and Russian-language memes and control stimuli were identified using instrumental neuroimaging methods. Memes constitute a very particular cross-cultural phenomenon, which is a combination of textual and illustrative information, and their effect on the functional state of the brain appears to be little studied. By demonstrating the textual and illustrative part of the memes separately and registering the EEG, we discovered how the text modulates the subsequent perception of the drawing by activating the mechanisms of visual attention. Reading the text in the native language caused a greater response of theta activity in the associative sensory areas of the cortex, which is associated with a better understanding of the meaning of the text. At the same time, the perception of illustrations to English-language memes caused a greater response of theta and alpha rhythms in most of the considered areas of the cortex, which reflects the processes of memory, emotional reaction and the involvement of large neuronal resources for the integration and understanding of the whole image of the meme.

Keywords: memes · EEG · attention · modulation · context · alpha-rhythm

1 Introduction

Modern memes play an important role in cultural communication and social dynamics. They are a combination of textual and visual information, causing

A. Karpov and V. Delić (Eds.): SPECOM 2024, LNAI 15299, pp. 201–215, 2025.
https://doi.org/10.1007/978-3-031-77961-9_15

complex cognitive and emotional reactions. We define "memetics as an interdisciplinary knowledge domain including, as a study object, methods for transmitting network information with the use of concise monocode or polycode ministructures characterized by maximum network virality and popularity" [1, p. 80]. In the process of analyzing a meme object, the following stages are distinguished: theme, types of meme encoding: objects, actions, etc., the modality of the meme: neutral, positive, or negative. Currently, all means of information transmission are focused on the developments in the field of Digital Humanities, which is primarily related to the process of communication, the transmission of verbal information in various languages of the world and the development of novel ways to enhance this process, taking into account the discursive-network basis of information transmission. Digital Humanities has a number of additional qualities of verbal information transmission, including speed, monocode/polycode representation of this information, and the utmost conciseness of memetic images. Social network discourse (SND) in memetics contributes to the development of thematic diversity, imagery and multi-layered information structure of the transmitted message.

From our point of view, memetics completely coincides with the functions of social network discourse (SND) in the broad meaning of this concept. It is worth mentioning the main distinctive features of SND, which are also important for understanding the specifics of memetics in digital communication. The identification of verbal and paraverbal specifics of the formation and functioning of SND in the global electronic media environment is based on its definition as a special electronic macropolylogue, taking into account the following types of categories of form, content and functional weight [2,3]: a) electronic macropolylogue—SND in form: distant; mediated; in real time (online) and deferred (offline); single—vector - polyvector; monochronous—polychronous; b) electronic macropolylogue—SND in content: monothematic—polythematic; information—rich (highly contextual)—information is not saturated (low-contextual); provoking controversy, specific actions, deeds - not provoking controversy, specific actions, deeds; c) electronic macro—polylogue - SND by function: informing, containing the point of view of the sender of the message; influencing, containing special linguistic means of influencing the recipient of the message; encouraging with a certain target attitude to commit specific actions, deeds (in particular, destructive ones, implemented according to the scheme "incentive → pragmatic reaction in the form of a specific destructive action"), manipulating the recipient's consciousness; designed for a target limited group of users—for an unlimited number of users; d) electronic macropolylogue—SND for accounting for influence factors on the specifics of communication: psychological and physiological (for example, age, gender, pathological, emotional, etc.); ethnic; socio-economic; political and geopolitical; confessional; cultural; pragmatic; moral and ethical.

Currently, memetics is analyzed mainly in connection with social and political topics, for example, in the works [4–11], etc. In the above-mentioned studies, the content dominant is polycode memes, which include various manifestations of

the real political life of a particular country. It should be emphasized that the obtained data imply the relevance of the meme as a means of pedagogical and sociological observation [12].

In this regard, we also studied that cognitive and neurophysiological re-coding of the processes in the brain reinforced by the constant and long-term use of the same foreign language stimuli-patterns, which leads to a change in the behavioral reactions of Internet users in the process of virtual network communication, as well as real communication [13]. Previously, we also collected a large-scale multimodal polycode linguistic database of memes using Big Data processing technologies and a deep annotation system for polycode texts [14].

The perception and processing of visual information in the human brain includes several stages, from the reception of light signals by the eye to their interpretation in the cerebral cortex [15–17]. The textual context presented together with the visual stimulus can significantly modulate the perception of this stimulus due to a number of cognitive and neurophysiological processes. This effect is based on the integration of various sensory modalities and contextual information that occurs in the brain. The textual context can activate relevant schemas and expectations associated with the information read in memory. Context can also change the activation of associative visual areas of the brain due to attention processes. Thus, by registering the EEG when perceiving the consistently presented text and illustration, it is possible to assess the induced changes in the electrical activity of certain areas of the cerebral cortex when reading the text and visually perceiving the illustration. Comparing evoked responses (event responses) during the perception of a simple description of an illustration with those observed in the case of an allegorical one (as in the case of memes), it is possible to identify neurophysiological correlates of contextual modulation of perception [18]. The notion of the induced changes in electrical activity suggests that we are investigating changes associated with the occurrence of some event—a stimulus or a task [18,19]. Desynchronization or suppression of alpha-range rhythms (ERD, Event-Related Desynchronization) is associated with activation of the visual cortex, processing and analysis of visual stimuli, cognitive load, increased visual attention [20,21]. The opposite effect, an increase in rhythmic activity, is called Event-Related Synchronization (ERS, Event-Related Synchronization) [19]. In this case, the increase in amplitude is probably mediated by the joint or synchronized behavior of a large number of neurons. In works with visual stimuli, synchronization occurs more often in the theta frequency range, and at the first moments of time is associated with the processes of involuntary attention [22], and then it can occur due to the inclusion of mechanisms of memorization or emotional response to the stimulus [23,24]. Neurophysiological studies of the perception of memes are still relatively rare, since the memes themselves are a relatively new but rapidly developing phenomenon. Thus, the EEG study of textual modulation of perception of visual stimuli using examples of memes in two linguistic paradigms is important for understanding the neuropsychological mechanisms of perception and integration of multisensory information, as well as for identifying cultural and linguistic dif-

ferences in cognitive and emotional reactions. This can have wide applications in the fields of cognitive neuroscience, psychology, marketing and intercultural communication.

This pilot study's primary aim was to identify key neurophysiological processes related to the perception of memes in Russian and English by native speakers of the Russian language using the electroencephalography (EEG) method. The proposed technique allowed us to analyze the key characteristics associated with the processing of multimodal stimuli and understand how the linguistic context modulates the perception of visual information.

2 Method

2.1 Stimuli Corpus

Selection of Memes. The first stage of this pilot study was the selection of stimuli. The search across the Internet was focused on the memes of a particular format: those containing text that defined a new context complementing and changing the perception of the graphic illustration, and the illustration itself. One of the requirements that determined the choice of the memes was that it should be possible to show the text and the illustration to it on the screen not simultaneously, but sequentially. This was necessary in order to separate the perception of the text and the viewing of the illustration in time—this way we could accurately synchronize the registration of the EEG with the reading of the text and with the perception of the picture. Another requirement was that the text itself without an illustration should not be self-sufficient. The illustration was supposed to fully reveal the theme of the joke and complement the context created by the text. The text or the illustration alone should not constitute a self-contained joke. The third important condition for the selection of memes to be more uniform is that the illustration should consist of one picture (cartoons, for example, were excluded). A total of 20 memes in English and 20 memes in Russian meeting the conditions and requirements outlined above were selected for the pilot experiment.

Reference Stimuli Selection. Alongside the main collection of stimuli, 40 reference stimuli (20 stimuli with a Russian contextual description and 20 stimuli with an English description) were selected. These stimuli contained images of the same kind that are typically used to create memes, but the text added to them did not create additional context, but only described the content of the picture.

2.2 Creation of Experimental Paradigm

As it was mentioned above, memes were shown not in the traditional text + picture format, but sequentially—first the text, and then the picture. This is necessary in order to separate the perception of the text and the illustration

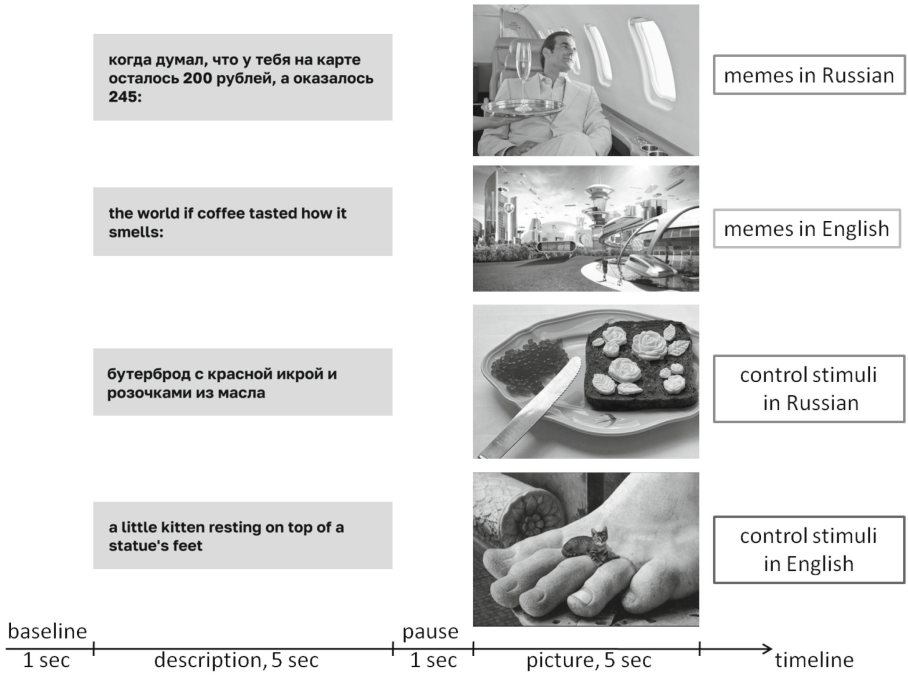


Fig. 1. The scheme of the experiment (the sequence of stimuli was randomized)

in time and register the EEG during the reading of the text and during the perception of the illustration.

The experimental paradigm included 20 meme stimuli in Russian, 20 meme stimuli in English, and the same numbers of reference stimuli in Russian and English. The design of the experiment and demonstration of stimuli was carried out using the Presentation software by Neurobehavioral Systems Inc. First, calibration samples were recorded at rest with eyes closed and open, and then the display of visual stimuli began in a pseudo-random order. Then the text component of the meme was shown on a gray background for 5 s, and after a 1-s pause the illustration of the meme was also shown for 5 s. After that, there was a pause between stimuli lasting 3 to 4 s. In the case of the control stimuli, a contextual description of the illustration was also presented first for 5 s, and then, after a 1-s pause, the picture itself was also shown for 5 sec (see Fig. 1).

2.3 EEG Registration

The pilot study was involved one subject without any neuropsychiatric disorders (female, 25 years old, English language proficiency level B2-C1). The 32-channel EEG was recorded using a BrainAmp DC amplifier (Brain Products GmbH, Germany) and Ag/AgCl electrodes arranged in accordance with the international 10–20 system (referent in position Fz). The data were recorded at a sampling

rate of 512 Hz; the impedance was maintained below 15 kOhm; a low-pass filter of 70 Hz, a high-pass filter of 1 Hz and a 50 Hz notch filter were used.

The registered EEG fragments were processed using the MNE-Python batch software. First, filtering was performed to remove too high and too low frequencies (the range of 2–40 Hz was left), and then the EEG signal was cleaned from artifacts by the method of independent ICA components. The independent components of the signal were calculated using the Infomax algorithm, topograms reflecting the localization of the component on the scalp model were calculated using the coefficients of the demixing matrix for each component. After detecting the artifact components, which are well detected due to their properties, we reset them. After that, we performed the reverse decomposition of the components into an EEG signal and already received an artifact-free recording. This method is widely used in modern EEG research.

2.4 Wavelet Transform and Analysis of ERD/ERS

After the preprocessing stage, the purified EEG was analyzed using the Morlaix wavelet transform, and the calculation of the ERD and ERS curves was performed. The signals of the selected components were sliced into fragments according to the marks of the beginning of the stimulus display. A 1-sec time fragment was taken as the baseline (background), immediately before the stimulus was given. In total, stimuli of 4 categories were presented to the subjects: Stimuli:

- Memes in Russian
- Memes in English
- Reference stimuli in Russian
- Reference stimuli in English

According to this categorization of stimuli, three-dimensional wavelet maps were averaged separately for each electrode. To obtain ERD/ERS curves, the MNE Python software package averaged three-dimensional wavelet maps along the time axis in four standard frequency ranges – delta rhythm (2–4 Hz), theta rhythm (4–8 Hz), alpha rhythm (8–13 Hz), lower beta rhythm (13–24 Hz). Further, the curves averaged individual areas of the cerebral cortex —'Occipital-parietal', 'Central-parietal', 'Central', 'Frontal', 'Temporal'. Thus, we obtained a set of ERD/ERS curves for each category of stimuli, for selected frequency ranges and individual cortical regions.

3 Results

The figures display averaged ERD/ERS curves for 20 similar stimuli in each stimulus category, based on the wavelet transform of EEG fragments recorded during their presentation. The line represents the average value at each point in time, with the spread (standard deviation) shown around the curve at each time point.

3.1 ERDS Curves in Text Perception

Differences in ERDS curves for different language categories of stimuli in the theta range in the central parietal and central regions of the cerebral cortex were observed in the first 1.5 s of text perception (see Fig. 2). Interestingly, reading the text in native Russian led to a greater response than reading the English text. The central parietal regions of the brain are the integrating zones of sensory stimuli, and there is also a Wernicke zone responsible for understanding speech. A larger amplitude response of theta activity to a native Russian-language text is most likely associated with a better understanding of its meaning, its figurative representation in the associative sensory areas of the cerebral cortex responsible for the integration of sensory information. Thus, the different response at theta frequencies in these areas reflects the processes of processing and understanding text in native and foreign languages.

We also obtained differences in the dynamics of the alpha rhythm in the occipital, parietal, parietal-central, frontal and temporal regions of the cerebral cortex when reading memes compared with descriptions of control stimuli (see Fig. 3). At 0.5–1 s into the demonstration the perception of meme texts caused desynchronization (suppression) of the alpha rhythm, and the perception of the text of control stimuli caused, on the contrary, synchronization of the alpha rhythm in these areas. Suppression or desynchronization of the alpha rhythm always indicates visual processing, increased concentration of attention to the task, while synchronization of the alpha rhythm indicates a decrease in attention. The results obtained indicate that the perception of the textual part of memes causes a stronger concentration of attention and greater involvement of these areas of the cortex in the process of processing and analyzing visual information.

3.2 ERDS Curves During the Anticipation of Illustration

We also analyzed a one-second fragment of the pause between the submission of the text part and the illustration to identify a possible difference in the moment of waiting for the demonstration of the image after reading the text. We found that after reading the text component of the memes, unlike reference stimuli, an increased concentration of visual attention in the parietal-occipital regions remained in the pause before demonstrating the picture (see Fig. 4). This is clearly indicated by a lower level of desynchronization of the alpha rhythm on the ERDS curves in the parietal-occipital visual areas of the cerebral cortex, which can be interpreted as an increase of interest and anticipation of a subsequent visual stimulus.

3.3 ERDS Curves During the Perception of the Illustration

The same alpha rhythm in the parietal-occipital (visual) areas of the cortex shows that at the beginning of the demonstration of illustrations, desynchronization or suppression of the alpha rhythm occurs much faster in the case

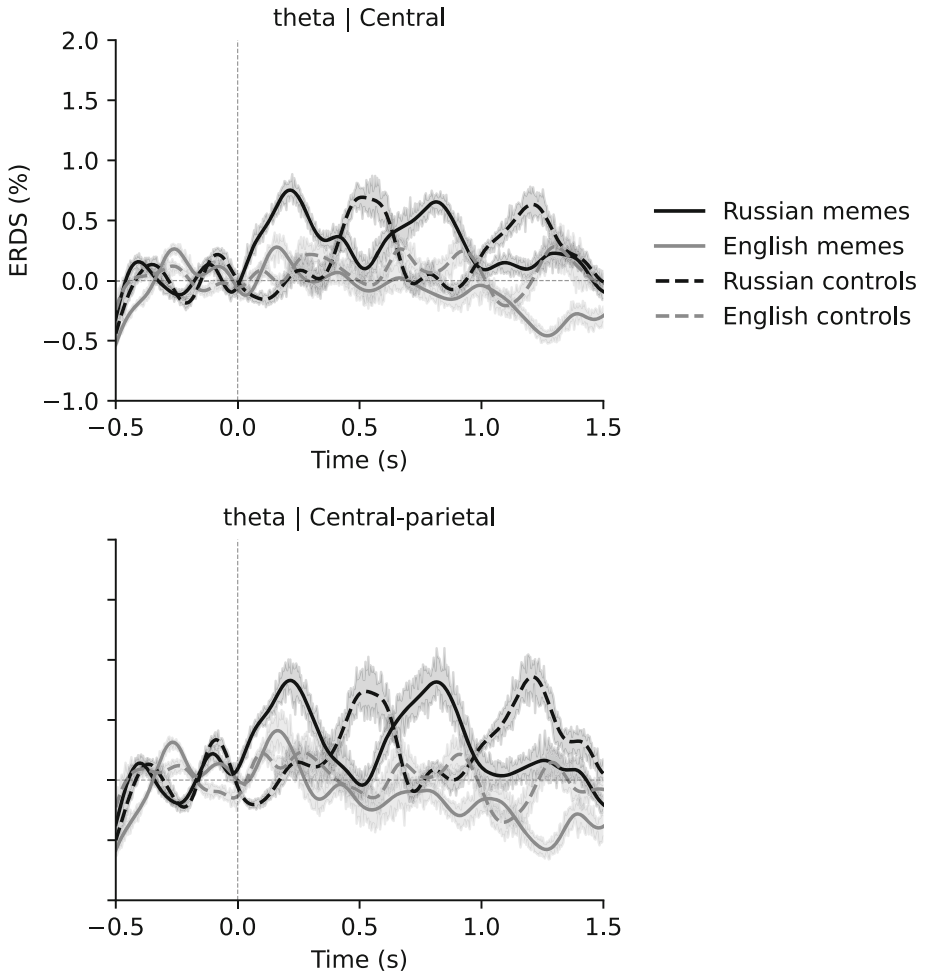


Fig. 2. Event-related synchronization in central and central-parietal cortical areas in theta-band during the perception of text in native and foreign languages

of memes. ERDS curves for control stimuli, on the contrary, show large synchronization peaks (see Fig. 5). Thus, reading the textual part modulates the functional state of the visual cortex and promotes its activation and retention of visual attention for faster perception of the illustration. It can also be noted that reading the textual part of memes arouses more interest in illustration, which appears not to be the case with the reference stimuli.

The following figures demonstrate the behavior of theta and alpha rhythms during the perception of illustrations at the initial stage (Fig. 6) and for the full 5 s (Fig. 7). It can be seen that theta activity was significantly higher at 0.5–1 s in the occipital, parietal and central regions of the cortex when

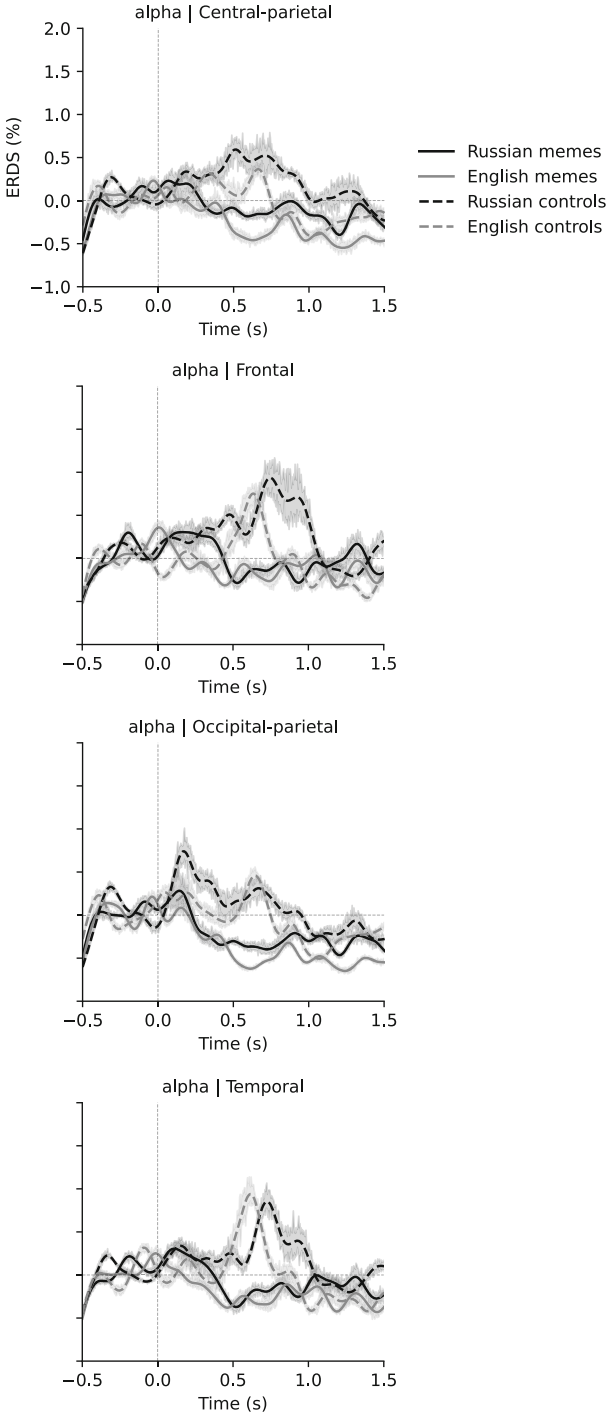


Fig. 3. Event-related desynchronization/synchronization in occipital-parietal, central-parietal, frontal and temporal cortical areas in alpha-band during the perception of the text component of memes/reference stimuli

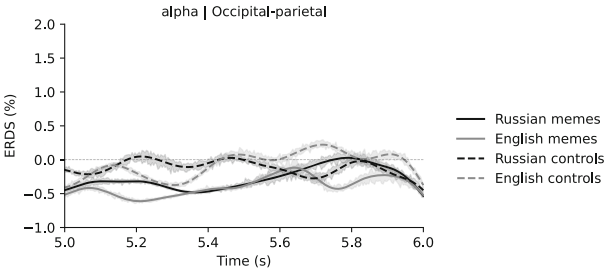


Fig. 4. Event-related desynchronization in occipital-parietal cortical areas in alpha-band while waiting for the demonstration of the illustration after reading the text of the memes

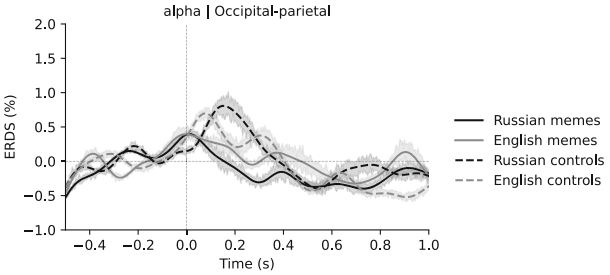


Fig. 5. Event-related desynchronization in occipital-parietal cortical areas in alpha-band at the moment of appearance of the illustration after reading the text of the memes

perceiving illustrations specifically for English-language memes. The theta rhythm is associated with various national brain processes, primarily with the processes of memorization, emotional reaction, and attention processes. Here we see that the illustration of an English-speaking mother causes a stronger surge of theta activity, apparently associated either with an emotional reaction or with the process of reproducing the text in English in memory. The following figure shows the dynamics of alpha activity in the central, frontal and temporal regions of the cerebral cortex. Here you can see that the strongest synchronization also occurs for English memes, and the weakest responses occur for categories of stimuli in native Russian. In general, it was expectable that stimuli related to a foreign language could elicit a stronger response, since they require more resources—attention, memory—for their processing. But it is interesting that we received a "delayed" enhanced response of rhythms not to the perception of a foreign text itself, but to the perception of an image associated with a foreign description. This may be another example of contextual modulation of the perception of an illustration.

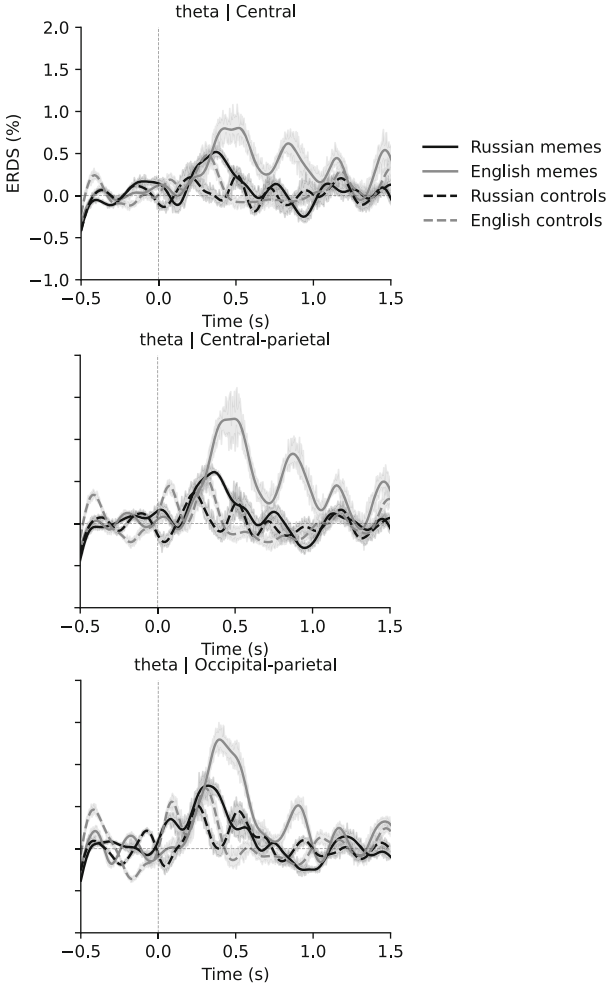


Fig. 6. Event-related synchronization in occipital-parietal, central-parietal and central cortical areas in theta-band at the moment of appearance of the image component of English language memes

3.4 Limitations

Conducting research on a single participant has significant limitations, primarily due to the high variability in individual reactions. Although this study does not answer the question of how consistently these differences would manifest across a larger sample, it does allow us to observe certain patterns and mechanisms in one individual. These patterns can later be confirmed or refuted in a larger experiment. Thus, this pilot study was designed to test the experimental methodology and identify preliminary findings that will help us formulate hypotheses

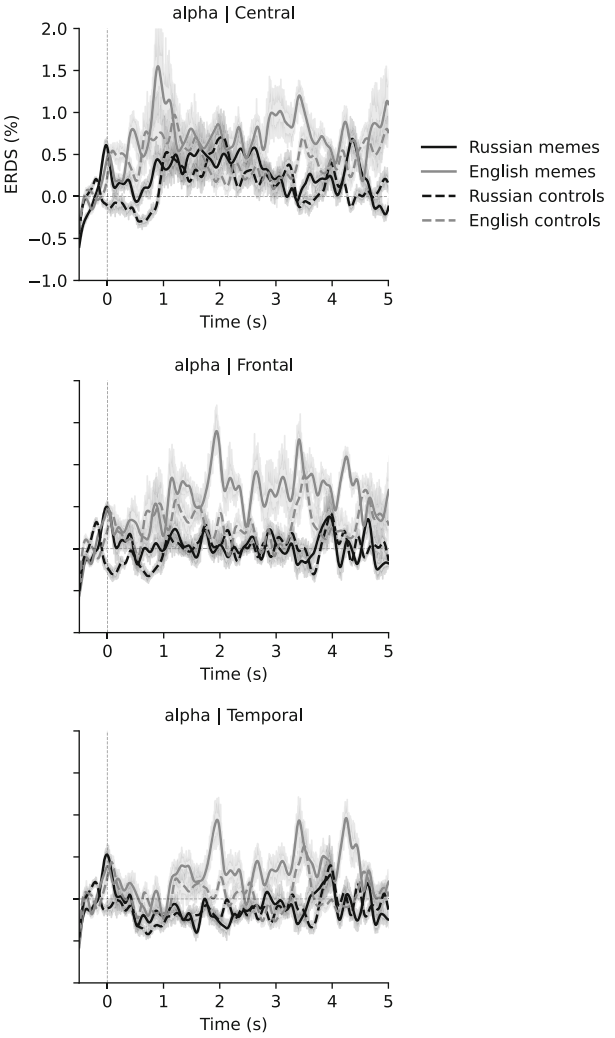


Fig. 7. Event-related synchronization in central, frontal and temporal cortical areas in alpha-band during the perception of the image component of English language memes and reference stimuli

for a full-scale study, while also providing early insights into how illustrations modulate text perception.

4 Conclusion

This pilot study aimed to investigate the neurophysiological correlates of how textual modulation affects the perception of visual stimuli, using Russian and

English as examples. To separate the perception of text and image and synchronize it with EEG data, the text and image of each meme were presented sequentially rather than simultaneously. The study also included reference stimuli consisting of similar images used in memes, but with descriptive text that merely explained the image. This allowed for a comparison between the effect of meme text, which creates a new context for the image, and descriptive text.

The findings revealed that reading Russian text elicited a stronger theta rhythm response in the central-parietal brain regions, which is linked to better understanding and emotional engagement compared to English text. Additionally, reading memes in both languages led to alpha rhythm desynchronization, indicating heightened attention, while control stimuli with descriptive text caused alpha rhythm synchronization, reflecting reduced attention.

Meme texts also maintained increased attention during the pause before the image was shown, which was not observed with control stimuli. This modulation of cortical function and sustained attention resulted in an earlier onset of alpha rhythm desynchronization—signaling visual processing in the visual cortex—when meme illustrations appeared, compared to control stimuli.

The analysis of ERDS curves showed that the perception of illustrations in English memes caused the strongest theta rhythm synchronization, which is associated with memory processes and emotional responses to stimuli. This may be due to the additional cognitive effort required to process and retain phrases in a non-native language, as the perception of English meme illustrations demanded more cognitive resources and repeated access to working memory to integrate the full meme (text + image).

In conclusion, this pilot study identified neurophysiological correlates of how text modulates the perception of visual stimuli, highlighting language-related differences in attention and memory processes. However, these results are preliminary and require validation with larger samples.

References

1. Potapova R., Potapov V.: Internet memetics as an emotiogenic environment of the network communication. Bull. Russ. Acad. Sci. Stud. Literature Lang. **81**(2), 78–91 (2022). (in Russian). https://doi.org/10.31857/S160578800019458-9
2. Potapova, R.: From deprivation to aggression: verbal and non-verbal social network communication In: Global Science and Innovation. Materials of the VI International Scientific Conference, Chicago, 2015, vol. 1., pp. 129–137 (2015)
3. Potapova, R.: Deprivation as a basic mechanism of verbal and paraverbal human behavior (based on social network communication) In: Potapova, R.K. (ed.) Speech Communication in Information Space, Lenand, Moscow., pp. 17–36 (2017). (in Russian)
4. Bown, A., Bristow, D. (eds.): Post Memes: Seizing the Memes of Production. Punctum books, Brooklyn (2019)
5. Egner, M.: Humor im Internet Analyse humoristischer Formen der Kommunikation in sozialen Medien. Masterarbeit. Universitaet Salzburg (2018). https://eplus.uni-salzburg.at/obvusbhs/content/titleinfo/5015234/full.pdf

6. Milner, R.: The world made meme: Discourse and identity in participatory media. PhD diss. University of Kansas (2012). http://hdl.handle.net/1808/10256. Accessed 11 June 2023
7. Milner, R.: Pop polyvocality: internet memes, public participation, and the Occupy Wall Street movement. Int. J. Commun. **7**, 2357–2390 (2013). https://ijoc.org/index.php/ijoc/article/view/1949/1015
8. Nowotny, J., Reidy, J.: Memes - Formen und Folgen eines Internetphaenomens (2022). https://www.transcript-verlag.de/media/pdf/70/97/c5/oa9783839461242.pdf
9. Osterroth, A.: Das Internet-Meme als Sprache-Bild-Text. IMAGE. Zeitschrift fuer interdisziplinaere Bildwissenschaft **11**(2), 26–46 (2015)
10. Osterroth, A.: Sprache-Bild-Kommunikation in Imageboards - Das Internet-Meme als multimodaler Kommunikationsakt und politisches Aergernis. Universitaet Koblenz-Landau. Koblenz and Landau (2016). https://www.uni-koblenz-landau.de/de/landau/fb6/germanistik/mitarbeiter/wissenschaftliche-mitarbeiter/andreas-osterroth/ArtikelGAL
11. Segev, E., Nissenbaum, A., Stolero, N., Shifman, L.: Families and networks of internet memes: the relationship between cohesiveness, uniqueness, and quiddity concreteness. J. Comput.-Mediat. Commun. **20**(4), 417–433 (2015)
12. Reidel, L.: Eine konsumentenorientierte Betrachtung der Memetik. GRIN Verlag, Muenchen (2019)
13. Potapova, R., Potapov, V., Gorbunov, P.: The brain activity of the bilingual code-switching communication. In: Wen, S., Yang, C. (eds.) Biomedical and Computational Biology. BECB 2022. Lecture Notes in Computer Science, vol. 13637, pp. 274–281. Springer, Cham (2023) https://doi.org/10.1007/978-3-031-25191-7_22
14. Potapova, R., Potapov, V., Gorbunov, P.: On the experience of statistical processing of memes in Big Data format. In: Yang, X.S., Sherratt, S., Dey, N., Joshi, A. (eds.) Proceedings of Ninth International Congress on Informatics and Communication Technology (ICICT 2024, London). Lecture Notes in Networks and Systems, vol. 1014, pp. 297–304. Springer, Singapore (2024). https://doi.org/10.1007/978-981-97-3562-4_24
15. Skrandies, W.: Visual information processing: topography of brain electrical activity. Biol. Psychol. **40**(1-2), 1–15 (1995). https://doi.org/10.1016/0301-0511(95)05111-2. PMID: 7647172
16. Melcher, D., Morrone, M.C.: Nonretinotopic visual processing in the brain. Vis. Neurosci. **32**, E017 (2015). https://doi.org/10.1017/S095252381500019X. PMID: 26423219
17. Klymenko, V., Coggins, J.M.: Visual information processing of computed topographic electrical activity brain maps. J. Clin. Neurophysiol. **7**(4), 484–497 (1990). https://doi.org/10.1097/00004691-199010000-00004. PMID: 2262542
18. Graimann, B., Huggins, J., et al.: Visualization of significant ERD/ERS patterns in multichannel EEG and ECoG data. Clin. Neurophysiol. **113**(1), 43–47 (2002). https://doi.org/10.1016/S1388-2457(01)00697-6
19. Pfurtscheller, G.: Event-related synchronization (ERS): an electrophysiological correlate of cortical areas at rest. Electroencephalogr. Clin. Neurophysiol. **83**(1), 62–69 (1992). https://doi.org/10.1016/0013-4694(92)90133-3
20. Karimova, E., Ovakimian, A., Katermin, B.: Live vs video interaction: sensorimotor and visual cortical oscillations during action observation. Cerebral Cortex **34**(4), bhae168 (2024). https://doi.org/10.1093/cercor/bhae168
21. Bazanova, M.: Sovremennaja interpretacija al'fa-aktivnosti elektrojencefalogrammy. Uspehi fiziologicheskih nauk **40**(3), 32–53 (2009). (In Russian)

22. Gulyaeva, A., Karimova, E.: Concepts and approaches to the study of visual spatial attention. Neurosci. Behav. Physiol. **3**(53), 416–431 (2023). https://doi.org/10.1007/s11055-023-01440-6
23. Lebedeva, N., Karimova, E., Potapova, R., Potapov, V.: Comprehensive study of the functional state changes when perceiving media content of different modality. Zhurnal vysshej nervnoj dejatel'nosti **71**(1), 86–103 (2021). https://doi.org/10.31857/S004446772101007X. (in Russian)
24. Potapova, R., Potapov, V., Lebedeva, N., Karimova, E., Bobrov, N.: The influence of multimodal polycode Internet content on human brain activity. Lect. Notes Comput. Sci. **12335**, 412–423 (2020). https://doi.org/10.1007/978-3-030-60276-5_40

Speech Synthesis and Perception

End-to-End Speech Synthesis for the Serbian Language Based on Tacotron

Tijana Nosek[1(✉)], Siniša Suzić[1], Milan Sečujski[1], Vuk Stanojev[1], Darko Pekar[2], and Vlado Delić[1]

[1] Faculty of Technical Sciences, University of Novi Sad, Novi Sad, Serbia
tijana.nosek@uns.ac.rs
[2] AlfaNum Ltd., Novi Sad, Serbia

Abstract. End-to-end text-to-speech (TTS) systems allow for the generation of high-quality computer-generated speech without relying on expert-created modules. This paper outlines initial efforts to develop a Serbian end-to-end TTS system using the Tacotron architecture. Listening tests revealed that while Tacotron can produce natural-sounding synthesis when properly trained, it is prone to overfitting and requires extensive data to avoid frequent hallucinations and accent errors. The use of a vocoder proved to be crucial in overall speech quality. Although the level of Tacotron training is less critical, it still demonstrates easy overfitting with relatively small databases. Correct accents and the absence of artifacts and hallucinations are extremely important for listeners, and any issues in these areas result in significantly lower ratings. Despite being less expressive, a controllable standard DNN-based TTS with a standard front end receives better grades because it never hallucinates and rarely makes linguistic mistakes. Integrating expert knowledge from existing pipelines can further improve synthesis quality, especially in data-constrained scenarios.

Keywords: Speech Synthesis · Tacotron · Deep Neural Networks · Front End

1 Introduction

Text-to-speech (TTS), also known as speech synthesis, aims to generate natural, expressive, intelligible speech from text mimicking human speech patterns. TTS has broad applications in human communication and has been a long-standing research topic in natural language and speech processing, as well as in artificial intelligence [1]. Over the decades, TTS systems have evolved from concatenative synthesizers, via statistical parametric speech synthesis, to models based on deep neural networks (DNN) [2]. With the development of deep neural networks, TTS systems have evolved from CNN/RNN-based models to transformer-based models, from auto-regressive models to other generative models, from cascaded acoustic models/vocoders to fully end-to-end models [3].

Developing a human-like TTS system requires both signal processing and linguistic background knowledge. In an attempt to bypass the need for linguistic knowledge, TTS systems have moved to end-to-end models that can be trained from scratch on the

A. Karpov and V. Delić (Eds.): SPECOM 2024, LNAI 15299, pp. 219–229, 2025.
https://doi.org/10.1007/978-3-031-77961-9_16

paired data set of < text, speech >. Some end-to-end TTS models like WaveNet [4] and FastSpeech 2 [5] are developed to directly generate waveforms from text. Others, like Tacotron [6], are trained to simplify linguistic and acoustic features converting them into linear-spectrograms, while others like NaturalSpeech [3], Tacotron 2 [7], DeepVoice 3 [8], FastSpeech [9] and FastSpeech 2 [5], predict mel-spectrograms from characters/phonemes. These models are augmented with a neural vocoder to generate waveforms.

End-to-end models do not require alignment information between text and speech and can be scaled with large amounts of acoustic data with transcripts. It can also be easier to adapt the model to new data. End-to-end models can be more robust than models that have separate components for text analysis front-end, acoustic model and vocoder since each component's errors can propagate [6]. Even though these models can produce state-of-the-art results, they suffer from slow training and inference speed, as well as necessity for large amount of high-quality speech corpus required for training, which proves problematic for low resource languages [10].

To the best of the authors' knowledge the results described in this paper present the first attempt to create an end-to-end TTS system in Serbian. There have been attempts in creating end-to-end systems in other South Slavic languages such as Macedonian [11, 12]. The system described in this paper is based on Tacotron 2 architecture. Since there are no datasets in Serbian large enough to enable training the model from scratch, the English model has been adapted using the Serbian speech dataset. To overcome the problem in generation of Serbian accented vowels, the authors propose the usage of previously developed expert based modules for accent prediction in Serbian [13].

The remainder of this paper is structured as follows: in Sect. 2 we will present key components of the model architecture and challenges that occurred during the training; in Sect. 3 we will present the results of subjective tests that have been performed for system evaluation, and in Sect. 4, we will discuss the results we obtained. We will give concluding remarks in Sect. 5.

2 Models and Approaches

In this section an overview of different models used in experiments will be given, as well as the description of data used for creating TTS voices.

2.1 Tacotron

Original Tacotron-2 architecture [7] consists of 2 modules: a recurrent sequence-to-sequence network with attention, which is used for predicting mel-spectrograms from an input character sequence, and a WaveNet [4] based vocoder, which generates time-domain waveform samples conditioned on the predicted mel-spectrograms. In all of our experiments WaveNet based vocoder is replaced by more efficient and better quality HiFi-GAN vocoder [14].

The mel-spectrogram predicting network consists of encoder and decoder with attention. The encoder consists of character-embedding layer, 3 convolution layers and bidirectional LSTM layer. The encoder output is passed through attention network and its output is further passed to an autoregressive decoder network producing mel-spectrograms as output. In our experiments we used the implementation presented in [15].

Since the training of Tacotron model is data intensive and there is not enough material in Serbian to train the model from scratch, the idea was to use Tacotron model already trained on LJSpeech dataset and adapt it to the Serbian database. The main change was the introduction of set of characters for Serbian language. We used Latin characters, with the exception of digraphs LJ, NJ and DŽ, which are conventionally treated as single letters, and replaced by Q, W and X respectively for convenience.

Although initial experiments showed promising results producing intelligible and good quality speech, we noticed its problems, most notably those related to generating appropriate accents. In order to mitigate these problems we extended the initial set of characters defined for Serbian to cover accents types representative for Serbian. The prediction of accents for the Serbian language was performed by the TTS front-end module, based on high-quality expert system using dictionaries and morpho-syntactic rules [13]. The description of accent used is given in Sect. 2.1.1.

We tried two different approaches for including accent information in system training. In first one a digit was added after each vowel to indicate a certain accent type or the absence of accent (e.g. *točak* would be represented as *to2ča0k*). In the second approach each accented vowel was presented by a different diacritic, (e.g. *točak* would be represented as *tòčak*). More details about accent types in Serbian are given in the following section.

2.2 A Note on Serbian Orthography

The Serbian language exhibits almost ideal phonemic orthography i.e. an orthography in which the graphemes correspond consistently to the phonemes of the language. An ideal correspondence between graphemes and phonemes would imply that each word is pronounced exactly as it is written, and hence that in a text-to-speech system explicit grapheme-to-phoneme conversion methods, based on dictionaries and/or conversion rules, are largely unnecessary, since the spelling of a word unambiguously and transparently indicates its pronunciation. However, neither of the two alphabets used for Serbian (Cyrillic and Latin) distinguishes between short and long vowels or rising and falling tones in Serbian, which is why a written vowel character (e.g. “e”) can stand for any of the 6 possible cases – a non-stressed short vowel (/e/), a stressed vowel with short falling accent: (/ȅ/), short rising accent (/è/), long falling accent (/ê/), a long rising accent (/é/), as well as post-accent long vowel (/ē/). A difference between the accents can imply a difference between word meanings, which is why pitch accents should be considered as relevant to the phonemic inventory. Marking differently accented vowels in the text (with digit suffixes from 0 to 5 or with different diacritics) can be compared to the use of explicit phonetic transcriptions in TTS systems for languages with non-phonemic orthography, and in this research it was carried out in order to help the system establish relationships between words and their pronunciations more easily under conditions of data sparsity.

2.3 Standard TTS with Neural Vocoder

Standard Serbian TTS consists of three blocks: front-end, which performs text normalization and produces a set of linguistic features, a DNN based block, which predicts some acoustic features using linguistic features as inputs, and a vocoder. The initial system based on the usage of deterministic WORLD vocoder is introduced in [16], while the system using neural HiFi-GAN vocoder is presented in [17].

The DNN block for acoustic feature prediction consists of two neural networks [16], one for prediction of phoneme durations and the other which predicts vocoder features based on input linguistic features and outputs of duration prediction network. Both networks consist of 3 feed-forward layers and one LSTM layer. This block was further improved by enabling multi-speaker training and applying target speaker adaptation as presented in [18].

2.4 HiFi-GAN Vocoder

A HiFi-GAN vocoder initially presented in [14] is a neural vocoder based on generative adversarial networks (GAN) [19]. A generative adversarial network typically comprises two main components: a discriminator and a generator. The generator produces data that mimics the statistical properties of the training dataset, while the discriminator's role is to determine whether a given sample is real or synthetic. HiFi-GAN, however, includes one generator and two types of discriminators. The generator in HiFi-GAN is a fully convolutional network that utilizes transposed convolutions and takes mel-spectrograms as input. The multi-period discriminator (MPD) consists of several sub-discriminators, each processing equidistant samples from the input speech, i.e. operating on a different sampling interval. This design allows the MPD to identify periodic patterns in the speech, working under the assumption that speech can be decomposed into sinusoidal components. Meanwhile, the multi-scale discriminator (MSD) analyzes consecutive samples from the input speech.

The process of adapting HiFi-GAN vocoder to standard Serbian TTS is described in [17]. The model is adapted from universal HiFi-GAN model trained on English data. This model was not trained directly on spectrograms extracted from natural speech but on data produced by specific guided acoustic network. In this way the model is better adapted to the outputs of a standard Serbian TTS system.

For the purposes of Tacotron based system the corresponding vocoder was also trained (finetuned). This vocoder was trained on mel-spectrograms produced by Tacotron by using text from original training dataset as Tacotron input. The target samples represent natural speech.

2.5 Training

All systems presented in the following subsections were trained using a Serbian speech corpus of a single female voice talent. This corpus was recorded in a professional studio and contains around 1.5 h of speech (including silent segments within utterances).

For the purposes of Tacotron training we used the same parameters as presented in the implementation given in [15], while the HiFi-GAN vocoders were trained using same hyper-parameter values given by the authors of original paper [14].

Both Tacotron and HiFI-GAN models were adapted by using starting models which were trained on LJSpeech dataset [20], which contains approximately 24 h of speech in English.

3 Experiments

For the evaluation and comparison of the selected models, several listening tests were performed, which will be presented in detail in following subsections. In each test 20 native Serbian speakers were included. Participants were instructed to use headphones to clearly hear even subtle differences in synthesized speech. None of the sentences used in tests were seen during the training of the models.

3.1 MUSHRA Test

The Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) test consists of 10 sets of utterances. Each set contains five utterances for grading and a clearly marked reference utterance. All utterances have the same linguistic content. The reference utterance contains natural speech of the target speaker. Among the five utterances for grading, one is identical to the reference utterance (hidden reference), while the other four are synthesized using different synthesizers. One of these four synthesized utterances is generated using the standard TTS model described in Sect. 2.2 (referred to as *st_TTS*), while the other three are synthesized by models based on Tacotron, described in Sect. 2.1. The first one generated by the model trained with a database not containing annotated accents (referred to as *TAC_noAcc*). The second one is the output of the model trained with accent information carried by a digit suffix, detailed in Sect. 2.1 (referred to as *TAC_Acc*), and the third one is the output of the model trained with accent information introduced through different diacritics (i.e. different characters) for each accented vowel, detailed in Sect. 2.1 (referred to as *TAC_Acc1*).

Listeners were asked to grade each of the five utterances by moving a slider on a scale from 0 to 100, allowing for very fine gradation of the quality of synthesized speech. The reference utterance served as an example of how natural speech should sound, and the same utterance was included among the five utterances for grading to verify if the listeners could identify and correctly rate it with a score of 100 or close enough.

The results (Fig. 1) showed that the reference utterance was graded almost 100, with an average score of 94.5. The lowest grade was given to *TAC_noAcc* (41.3), followed by *TAC_Acc* (50.9). The *st_TTS* and TAC_Acc1 received much better grades, with average scores of 64.8 and 69.6, respectively.

3.2 MOS Test

The Mean Opinion Score (MOS) test consists of 18 utterances with different linguistic content. One third of the utterances are synthesized with *st_TTS*, another third with *TAC_noAcc*, and the rest with *TAC_Acc1*. Among the six utterances produced by *TAC_noAcc*, half of them contain at least one incorrectly accented word, while in the rest all words are correctly accented. Among the six utterances produced by *TAC_Acc1*,

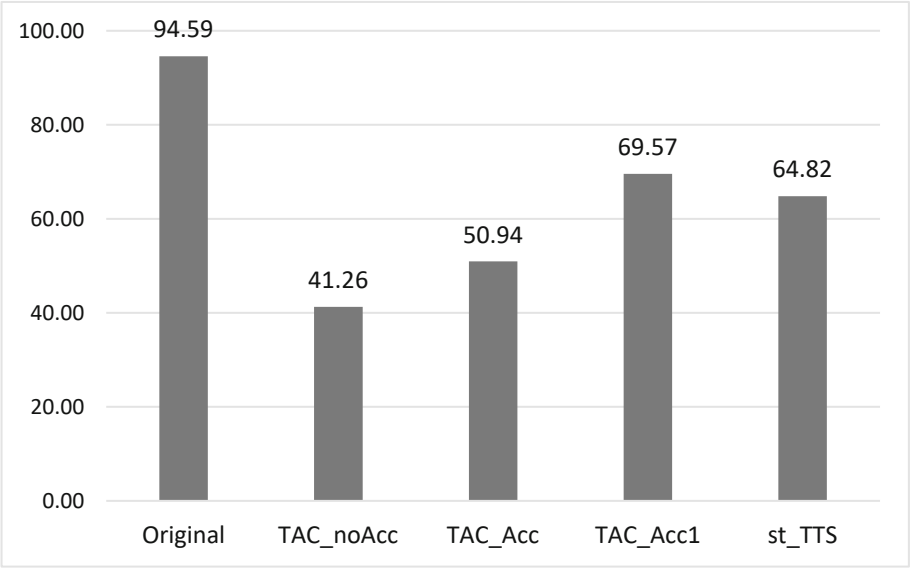


Fig. 1. Results of MUSHRA test – grades of 1–100 scale for quality of different synthesizers.

half of them contain hallucinations at the end of the utterance, while the rest are free of hallucinations (randomly synthesized non-existent phonemes).

Listeners were asked to grade each utterance on a 1–5 scale in terms of speech quality, i.e., its naturalness and intelligibility. A grade of 1 indicates unnatural and/or unintelligible speech.

The *st_TTS* received the highest average grade, 4.7, followed by *TAC_Acc1* with 4.1, and *TAC_noAcc* received the lowest grade, 3.0 (Fig. 2). However, when graded separately, the utterances produced by *TAC_Acc1* without hallucinations had an average grade of 4.6, almost as high as *st_TTS*, while those with hallucinations were graded 3.5 on average. Similarly, the utterances produced by *TAC_noAcc* with correctly accented words had an average grade of 3.6, while those with incorrect accents were graded 2.5 on average. The presence of hallucinations and incorrect accents in synthesized speech significantly lowered the perceived quality, resulting in grades lower by over 1 point.

3.3 Preference Test

In the preference test, there were 14 pairs of utterances. Each pair contained two utterances with the same linguistic content but produced by different synthesizers. All utterances are produced by models based on Tacotron. Eight pairs of utterances were used to analyze the impact of training the Tacotron model for different numbers of epochs, while the remaining pairs focused on the importance of adapting the HFG-based vocoder to the target speaker. In the first eight pairs, one utterance was produced by a less-trained model, while the other was produced by a more-trained model, with both utterances in each pair produced by the same vocoder. Two out of the eight pairs were produced by models trained with accent information, with one model trained for 250 epochs and the

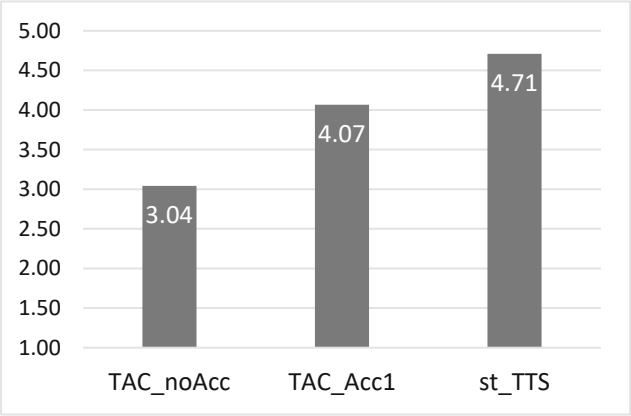


Fig. 2. Results of MOS test – grades for 1–5 scale for quality of different synthesizers.

other for 100 epochs. The rest of the pairs were produced by models trained without accent information, trained for 100, 300, 500, and 900 epochs. Each pair of models was compared. In the last six pairs of utterances, each pair contained one utterance produced with a universal model of HFG and the other with an HFG model trained for the specific Tacotron model used in both utterances.

Listeners are asked to choose the better, i.e. the more natural sounding utterance between the two in each pair, but they are also allowed to choose "no preference" as well.

The results presented in Fig. 3, show that listeners slightly prefer utterances generated by the Tacotron model trained for a longer time. There is also preference in favor of using HFG model adapted to target speaker compared to using universal HFG model as show in Fig. 4. However, in either case the differences are not significant.

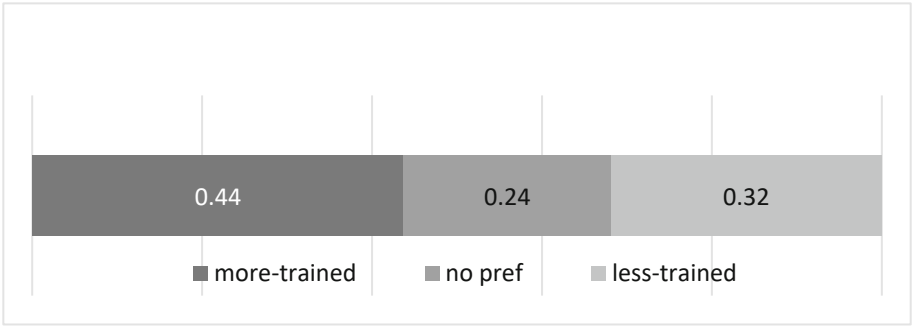


Fig. 3. Results of preference test – more or less trained Tacotron models.

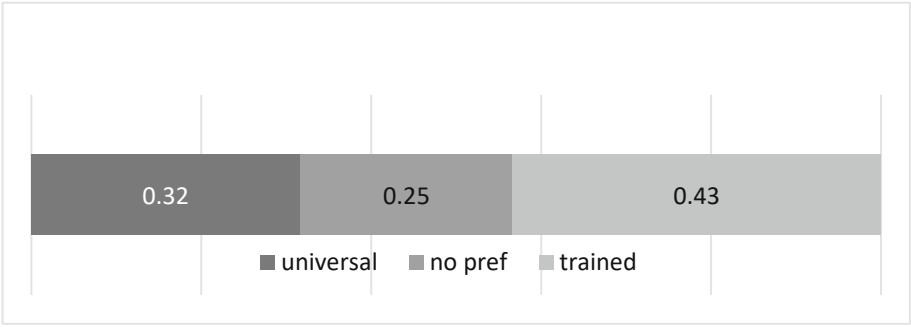


Fig. 4. Results of preference test – universal vs trained HFG model.

4 Discussion

The listening tests provided clear insights into how people perceive the quality of different speech synthesizers and what their main objections are. The MOS test showed that people perceive standard TTS as being of very high quality, grading it 4.7 out of 5 on average. Tacotron-based synthesizers received significantly lower grades, but a more detailed analysis reveals some key conclusions.

Firstly, the differences in grades between the two Tacotron-based models (3.0 and 4.1) indicate that training the same model with and without explicit information about accents in Serbian is crucial for improving the model. This is likely due to the system's inability to properly handle ambiguous vowel characters without sufficient data. When comparing synthesis from the same model trained without accent information, it received a grade of 3.6 for utterances with correct accents and 2.5 for utterances with incorrect accents. The presence of incorrect accents in Serbian not only impairs the naturalness of the synthesis but can also render speech unintelligible or change the meaning of an utterance. It is thus not surprising that the most significant objections from listeners are related to incorrect accents. This problem is largely mitigated by providing accent information during both training and synthesis.

Two methods for incorporating accent information were used: one involving adding accents as additional characters, so that combinations of subsequent characters (vowel + accent) provided full information. In the other approach we adopted, different characters were used for each possible vowel/accent combination, thus providing full information with just one character, although this increased data sparsity. To analyze performance, we synthesized 50 utterances with each of the three Tacotron-based models: the one without accent information (*TAC_noAcc*), the one with accents given as separate characters (*TAC_Acc*), and the third model with different characters for each vowel/accent combination (TAC_Acc1). *TAC_noAcc* produced utterances with at least one incorrectly accented word in 74% of utterances, *TAC_Acc* in 6%, and TAC_Acc1 only in 4% of all utterances. These results suggest that the proposed approaches utilizing accent predictions significantly reduce the problem of incorrect accents even with a relatively small training dataset.

Another problematic aspect of Tacotron-based models, especially when insufficient data is used for training, is the occurrence of hallucinations. These are manifested as randomly synthesized non-existent phonemes, usually at the end of an utterance, or by repeating the last phoneme from the input sentence. While hallucinations do not greatly impact overall intelligibility and naturalness, they are extremely annoying and negatively affect people's perception of the synthesizer's quality. By examining 50 utterances we conclude that hallucinations occur in 56%, 74% and 94% of them, in *TAC_Acc, TAC_noAcc* and TAC_Acc1, respectively. Additionally, in about 10% of utterances, the synthesis is completely unusable as the system fails to produce anything intelligible. The MOS test showed that people rated TAC_Acc1 synthesis at 4.6 when there were no hallucinations, but 3.5 when hallucinations were present. The hallucination problem can only be reduced by providing more training data in case of this model/architecture.

Although the *st_TTS* was graded as the best in the MOS test, likely due to the absence of any hallucinations and incorrect accents, owing to its front-end module and high controllability, listeners gave a slight advantage to *TAC_Acc1* in the MUSHRA test. A more detailed analysis of MUSHRA results shows that natural speech received a grade of 94.6, which is expected, while the next highest grade was 69.6. This significant gap indicates that synthesized speech is still easily distinguishable from natural speech, especially when directly compared with the same utterances produced by natural speakers. The lower grades for *TAC_Acc* and *TAC_noAcc* can be attributed to the more frequent occurrences of incorrect accents and hallucinations, as previously discussed.

However, the slightly lower grade for *st_TTS* compared to TAC_Acc1 (64.8 vs. 69.6) can be explained by the more lively or dynamic synthesis produced by the Tacotron-based model. Although *st_TTS*, when heard alone without any artifacts, hallucinations, or mistakes, sounds very good (receiving a grade of 4.7 out of 5 in the MOS test), hearing it together with Tacotron-based synthesis with the same linguistic content can highlight its lack of expressiveness (Fig. 5).

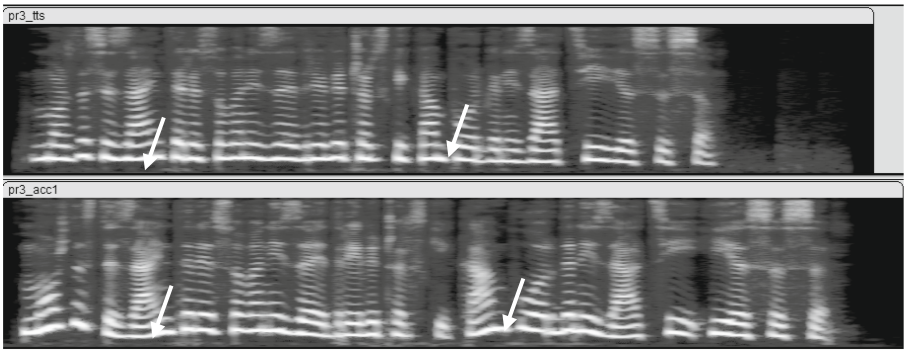


Fig. 5. Spectrograms of utterances with the same linguistic content produced by different synthesizers (the upper one produced by *st_TTS*, the lower one produced by *TAC_Acc1*).

Finally, as regards the results of the preference test, the lack of a clear difference between Tacotron models trained for more or fewer epochs can be explained by Tacotron's tendency to overfit easily, although the more trained versions were slightly

favored. Additionally, the authors confirmed that training for more epochs did not reduce the percentage of produced hallucinations nor did it improve accent learning.

Another conclusion from the preference test is that there is no significant difference between using a trained or universal HFG-based vocoder, although the trained one had a slight advantage. The authors find it more significant to use the trained version of the vocoder. The reason is the occurrence of artifacts and slight buzzing when using the universal HFG-based vocoder, but these issues were probably not prominent or annoying in the short and few examples that listeners heard during the test.

5 Conclusion

In this paper, we present a TTS (Text-to-Speech) system for end-to-end synthesis in Serbian, based on the Tacotron architecture. Due to the lack of a large, high-quality speech database in Serbian, the system was created by adapting a pre-trained English model. Initial experiments revealed issues with appropriately generating accents in Serbian. To address this, the authors proposed two methods involving modules for accent prediction from text. The approach using different symbols for each accented vowel produced better results. Although the Tacotron-based system can outperform the current best Serbian synthesizer, which uses separate front-end and DNNs, in some contexts, errors typical of sequence-to-sequence models, such as hallucinations and repetitions, significantly decrease the overall performance of the system.

Future work will include attempts to overcome data sparsity problems, especially with accents, by augmenting the training set using TTS-generated data. The authors also plan to explore newer architectures.

Acknowledgments. This research was supported by the Science Fund of the Republic of Serbia, Grant No. 7449, Multimodal multilingual human-machine speech communication, AI-SPEAK.

References

1. Tan, X., Qin, T., Soong, F., Liu, T.Y.: A survey on neural speech synthesis. arXiv preprint arXiv:2106.15561 (2021)
2. Delić, V., et al.: Speech technology progress based on new machine learning paradigm. Comput. Intell. Neurosci. **2019**(1), 4368036 (2019)
3. Tan, X., et al.: Naturalspeech: end-to-end text-to-speech synthesis with human-level quality. IEEE Trans. Pattern Anal. Mach. Intell. (2024)
4. Van Den Oord, A., et al.: Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499 12 (2016)
5. Ren, Y., et al.: Fastspeech 2: Fast and high-quality end-to-end text to speech. arXiv preprint arXiv:2006.04558 (2020)
6. Wang, Y., et al.: Tacotron: Towards end-to-end speech synthesis. arXiv preprint arXiv:1703.10135 (2017)
7. Shen, J., et al.: Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions. In: 2018 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP), pp. 4779–4783. IEEE (2018)

8. Ping, W., et al.: Deep voice 3: Scaling text-to-speech with convolutional sequence learning. arXiv preprint arXiv:1710.07654 (2017)
9. Ren, Y., et al.: Fastspeech: fast, robust and controllable text to speech. Adv. Neural Inf. Proc. Syst. **32** (2019)
10. Mu, Z., Yang, X., Dong, Y.: Review of end-to-end speech synthesis technology based on deep learning. arXiv preprint arXiv:2104.09995 (2021)
11. Mishev, K., Karovska Ristovska, A., Trajanov, D., Eftimov, T., Simjanoska, M.: MAKEDONKA: applied deep learning model for text-to-speech synthesis in Macedonian language. Appl. Sci. **10**(19), 6882 (2020)
12. Sofronievski, B., et al.: Macedonian speech synthesis for assistive technology applications. In: 2022 30th European Signal Processing Conference (EUSIPCO), pp. 1183–1187. IEEE (2022)
13. Secujski, M.S.: Obtaining prosodic information from text in Serbian language. In: EUROCON 2005-The International Conference on Computer as a Tool, vol. 2, pp. 1654–1657. IEEE (2005)
14. Kong, J., Kim, J., Bae, J.: Hifi-gan: generative adversarial networks for efficient and high fidelity speech synthesis. Adv. Neural. Inf. Process. Syst. **33**, 17022–17033 (2020)
15. NVIDIA. Tacotron 2. GitHub repository, https://github.com/NVIDIA/tacotron2. Accessed 23 May 2024
16. Delić, T., Sečujski, M., Suzić, S.: A review of Serbian parametric speech synthesis based on deep neural networks. Telfor J. **9**(1), 32–37 (2017)
17. Suzić, S., Pekar, D., Sečujski, M., Nosek, T., Delić, V.: HiFi-GAN based Text-to-Speech Synthesis in Serbian. In: 2022 30th European Signal Processing Conference (EUSIPCO), pp. 2231–2235. IEEE (2022)
18. Secujski, M., Pekar, D., Suzic, S., Smirnov, A., Nosek, T.V.: Speaker/style-dependent neural network speech synthesis based on speaker/style embedding. J. Univers. Comput. Sci. **26**(4), 434–453 (2020)
19. Goodfellow, I., et al.: Generative adversarial networks. Commun. ACM **63**(11), 139–144 (2020)
20. Keith Ito. The LJ Speech Dataset. https://keithito.com/LJ-Speech-Dataset. Accessed 23 July 2024

ChildTinyTalks (CTT): A Benchmark Dataset and Baseline for Expressive Child Speech Synthesis

Shaimaa Alwaisi(✉), Mohammed Salah Al-Radhi, and Géza Németh

Department of Telecommunication and Artificial Intelligence,
Budapest University of Technology and Economics, Budapest, Hungary
shaima.alwaisi@edu.bme.hu, {malradhi,nemeth}@tmit.bme.hu

Abstract. Designing expressive speech synthesis for child voice remains an unresolved problem. One of the major dilemmas faced by child TTS systems and child speech synthesis is the scarcity of datasets to train opaque data-hungry DNN-based models. Only a few datasets were proposed for the purpose of building child conversational AI agents, and many of them come with challenges such as noisy data and indiscernible speech. With this in mind, we introduce the ChildTinyTalks (CTT) dataset, comprising 2 h of speech collected from 25 kids in grades ranging from third to fourth grade, who are telling stories and sharing their experiences. The new dataset containing 1200 audio samples has been transcribed at the word level, comprising 4 classes of voice expressions. To verify the effectiveness of CTT in real-world situations, AutoVocoder models were trained and synthesized samples were generated. The models were trained on both the LJSpeech large scale dataset and our CTT dataset. Initial experimental results indicate that the CTT dataset can steadily give comparable results with acoustic model trained on a large-scale dataset with a size of less than 10% of the large dataset.

Keywords: Child Speech Dataset · Child Speech · AutoVocoder

1 Introduction

Despite the success stories in expressive Text-To-Speech TTS and speech synthesis models for adults [1–3], designing fully expressive TTS for children remains a challenge. One of the primary factors contributing to these successes is the use of high-quality benchmark datasets and low rate of errors in scientific challenges [4, 5]. Nevertheless, the limited availability of suitable training child datasets has hampered the development of TTS and Expressive TTS systems for children. However, it is not surprising that there is a lack of this kind of dataset, as collecting an appropriate amount of child speech involves many difficulties [6]. Typically, children are unable to articulate all speech units present in the language, they have limited reading skills and short attention spans. These issues result in irregular speech styles and imperfect recording samples [7].

The My Science Tutor MyST Dataset [8] can be considered the largest-scale available child speech corpus for research purposes. The dataset includes 393 h of conversational

A. Karpov and V. Delić (Eds.): SPECOM 2024, LNAI 15299, pp. 230–240, 2025.
https://doi.org/10.1007/978-3-031-77961-9_17

child speech, with a total of 228,874 utterances. Approximately,197 h of the dataset were transcribed. On the other hand, MyST comes with many problems, such as utterances without any phonetic meaning, noisy background. Moreover, there is poor audio quality, characterized by children speaking very close to the microphone. Through manual examination, some transcriptions were allocated to incorrect audio samples completely. As an illustration, within the supplied transcription, there was the sentence, "No, I don't hearing even a candle burns". Yet, in the actual transcription, it reads, "No, I don't hear anything when the candle burns".

In the domain of Automatic Speech Recognition (ASR) systems, it has been observed that even the robust models trained on adult data exhibit suboptimal performance when applied to child voices. This is attributed to the inherent differences between adult and child speech. Besides less developed speech production and perception models they have higher fundamental frequency, and shorter vocal tract length [9].

To bridge the performance gap in ASR models between adults and children, several well-known child datasets such as CMU Kids dataset [10], OGI Kids' Speech Corpus [11], Tball corpus [12], Providence Corpus [13] were used for training and testing ASR systems. However, Child ASR remains a challenging task due to various issues associated with these datasets. These issues include the absence of correct labels, instances of children not vocalizing, audio containing noise, difficulties in accessing the dataset [14].

To improve the performance of ASR and expressive speech synthesis models, we need to use high-quality child datasets for training or adapt the existing acoustic models. To this end, we propose the ChildTinyTalks (CTT) dataset. CTT stands as an expressive child dataset, encompassing 2 h of speech gathered from 25 students in grades spanning from third to fourth grade. These students engage in storytelling and recounting their experiences. The new dataset containing 1200 audio samples has been transcribed at the word level, comprising 4 classes of voice expressions. The source of the dataset is TEDxKid recordings [15].

Our paper is organized in the following way. In Sect. 2, we provide an overview of existing child datasets and popular techniques to address the problem of child dataset scarcity. Section 3 highlight our data collection and labelling. In this chapter, we also give a description and present key statistics of our dataset. Section 4 includes our baseline AutoVocoder speech synthesis model and a comparative study between the performance of the AutoVocoder trained on both our CTT dataset and the large-scale LJ Speech dataset [16]. In Sect. 5, we give the accuracy metrics and report the results of training our baseline model using both datasets across both objective and subjective testing benchmarks. Finally, Sect. 6 sums up our conclusions.

2 Related Work

The most attractive datasets for training child TTS and ASR models are: MyST, CMU Kids dataset [10], OGI Kids' Speech Corpus [11], Tball corpus [12], and Providence Corpus [13]. However, each of them comes with its own set of limitations. For instance, they lack expressive styles, which are crucial for training expressive child TTS models. Moreover, the absence of explicit correctness labels, audio samples without transcriptions and noisy data.

To fill this gap, many studies have aimed to tackle these challenges by adapting the acoustic features of child speech to align with those of models trained on adult speech [17]. Most of them used transfer learning approaches from adult to child speech and found it to be very helpful [18–20]. Maximum Likelihood Linear Regression (MLLR) [21], and stochastic feature mapping (SFM) [22] have been employed as adaptation methods, proving beneficial for adapting to child speech. The trend indicated that as the age of the child speaker increased, there was a reduced need for data adaptation. Consequently, younger children encountered more mismatch with adult speech. Other efforts have attempted data augmentation setups to increase training data by adding child speech. In general, this has not proven fruitful [23, 24].

For expressive child speech, a few studies have focused on exploring emotions in children's stories [25, 26]. To address the limited availability of child datasets, we introduce the CTT dataset. It has been transcribed at the word level, comprising four classes of expressive styles: sadness, excitement, happiness, and neutral.

3 ChildTinyTalks (CTT)

3.1 Description

The main idea of crawling speakers on YouTube was inspired by [27, 28]. They exclusively utilize YouTube as their primary source of data. CTT includes over 1200 utterances from 25 speakers extracted from TEDx Talks for Kids events uploaded to YouTube. We formulated a hypothesis suggesting that there are several YouTube channels, featuring children speaking. Typically, these channels are visually distinguishable, either by examining a grid of video previews or automatically through clustering sequences of audio speaker embeddings from videos. Based on this idea, we focused on TEDx Talks for Kids events, as this channel offers high-quality content recorded in a silent environment. Additionally, the speakers range in age from 6 to 11 years.

The key statistics of our dataset are described in Table 1. The gender distribution for CTT dataset is 52% and 48% for boys and girls respectively.

The language is only American English. The percentage of styles is 30% sad, 35% excited, 15% happiness and 20% neutral. Ten children from 6 to 7 years, fifteen children from 8 to 11 years. We intend to make CTT publicly available, but it requires agreements with the TEDx for Kids copyright holder, which are currently in progress.

3.2 Data Collecting and Filtering

The TEDx YouTube channel has been processed and filtered based on the number of available child videos, focusing on ages ranging from 6 to 11 years old. Audio was extracted from all TEDx for Kids videos that passed filtration phase.

3.3 Expressive Styles in ChildTinyTalks

CTT is a high-quality expressive speech dataset that includes both expressively rendered and presenting speech. The dataset covers four classes of expressive styles (sadness,

Table 1. ChildTinyTalk CTT description.

Dataset	Statistics
Number of POI	**25**
Number of Utterances	**1200**
Number of hours	**2**
Number of filtered Videos	**100**
Number of videos per POI	**3**
Avg. Number of utterances per POI	**48**
Avg. Length of utterances [s]	**6.71**

excitement, happiness, and neutral), so it can be used to build expressive child synthesis models. These four speaking styles are uttered by girls and boys in the English language. We offer illustrative audio samples from the CTT dataset and showcase vocoded samples achieved by baseline models are available online[1].

Table 2. Examples of captions in ChildTinyTalk CTT.

Style	Caption
Sadness	No one would invite him to their birthdays or include him in their group of friends it made both me and Ben really sad
Excitement	I feel good knowing I can do something for others to make them feel happy
Neutral	Number one think on the positives like I did in my room
Happiness	You're right it was amazing I even got, to do a campaign for red nose day it's, where everyone comes together to get rid of child poverty

3.4 Data Preprocessing

All audio files were decoded to and downsampled to mono wav format (sampling rate: 22.05 kHz, 16 bit PCM quantization) commonly used for speech and voice recognition tasks. We performed downsampling of the source signals of the audio channel of MP4 44.1 kHz in FLAC encoding. We entirely omitted any visual information from the TEDx channel. Sometimes, the audio files may contain substantial periods of silence, we removed all such silence regions, leaving only a small fraction at the beginning and end of the audio samples roughly 4 s. Audio samples with a significant amount of noise or crosstalk were removed with the help of Praat software [29]. The samples have been transcribed at the word level. We corrected obvious spelling errors in the transcriptions. We attempted to preserve explicitly mispronounced words to the greatest extent possible.

[1] https://github.com/shaimaalwaisi/ChildTinyTalks-CTT-dataset.

To ensure consistent Mel-spectrograms later in the training process, the lengths of audio signals are trimmed to averaging 5.5 s. To address scenarios where the audio signal is shorter than the desired segment size, it is padded with zeros to match the specified segment size. This guarantees uniform lengths for all audio signals presented to the neural model. For longer segments than 5.5 s, the excess portions are typically trimmed to ensure that all audio signals conform to the specified length. This approach maintains consistency across the dataset and ensures that the neural model processes inputs of uniform length.

4 Experiments

To demonstrate the efficiency of CTT in real-world situations, AutoVocoder models were trained and used for test waveform generation [30]. The baseline AutoVocoder [30] model was trained on both the LJ speech dataset and our CTT dataset. The comparison with the adult dataset, LJSpeech, was conducted due to the limited availability of comparable expressive child speech datasets. AutoVocoder is a voice encoder designed for training on speech waveforms with low computational cost.

We chose AutoVocoder due to its state-of-the-art capabilities in speech synthesis. It has demonstrated outstanding results, particularly in fine-tuning for child speech [31]. The encoder in AutoVocoder starts by converting time-domain signals into frequency-domain representations using a differentiable Short-Time Fourier Transform (STFT). From the resulting complex spectrum, four components magnitude, phase, real, and imaginary are extracted and treated as distinct channels, which are then processed by a convolutional residual network. This network consists of 11 basic blocks, each containing two 2D convolutional layers (kernel size 3), 2D batch normalization, and ReLU activation. Residual connections are applied to sum the input and output when the channel counts match. The first five blocks operate with four input and output channels, the middle block reduces the output to one channel, and the last five blocks maintain the single channel structure. A final linear layer compresses the dimensionality of each frame to match the size of a typical mel spectrogram frequency dimension. The decoder mirrors this process to reconstruct the waveform. Crucially, the architecture is non-autoregressive, allowing each frame to be processed independently, reducing computational load and speeding up waveform generation (Fig. 1).

The configuration of the AutoVocoder involved Adam optimizer, a Batch Size of 16, and Learning Rate set at 0.0002. Training was conducted on a server running Ubuntu 16.04.7 LTS, with NVIDIA-SMI and CUDA version 11.4 for efficient utilization of GPU type NVIDIA TITAN Xp. The audio samples in the dataset have been split into 80% for the training set and 20% for the testing set. The training set was used to train the Autovocoder on the features of both groups of children, The validation set from both age groups was used to evaluate and fine-tune the model's performance during training.

4.1 Objective Test

Mel-Cepstral Distortion MCD (dB): We utilized the mel-cepstral distortion (MCD) [32] as a distance metric to objectively evaluate the overall synthesized sound quality.

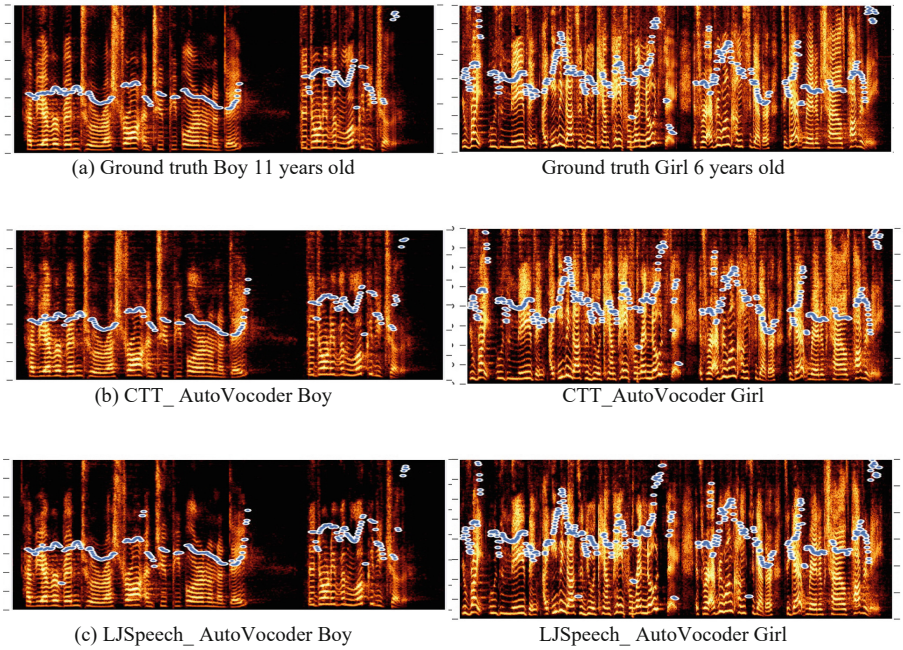


Fig. 1. Example of melSpectrogram and F0 comparison between reference and child audio synthesized for a boy and a girl:(a) ground truth spectrogram and F0, (b) CTT_AutoVocoder spectrogram and F0 Trained on our dataset CTT, and (c) LJSpeech_ AutoVocoder spectrogram and F0 trained on LJ speech 1.1 dataset. The horizontal axis gives the time dimension for the audio, while the left vertical axis represents the frequency dimensions. The right vertical axis represents the fundamental frequency.

It is commonly used to quantify the difference between two time-aligned mel-cepstral sequences. MCD is a common measure employed to quantify the dissimilarity between two time-aligned mel-cepstral sequences. For both models, we conducted an average MCD calculation across twelve synthesized sound samples, encompassing both girls and boys. The samples synthesized by AutoVocoder trained on the CTT dataset and the corresponding ground-truth samples, as shown in Table 2. The synthesized speech generated by the AutoVocoder closely aligns with the characteristics of the ground-truth reference speech. MCD values are calculated based on the following equation:

$$\text{MCD} = \frac{1}{N}\sum_{j=1}^{N}\sqrt{\sum_{i=1}^{D}\left(y_{i,j} - {}^{'}y_{i,j}\right)^2}, \tag{1}$$

where $\acute{y}_i$ represents the i^{th} coefficient of the generated MCCs, while y_i denotes the i^{th} coefficient of the ground truth MCCs.

The smaller the MCD between the synthesized and original mel-cepstral sequences, the higher the similarity between the synthesized and original speech. CTT_AutoVocoder, trained on the CTT dataset, exhibited comparable results to the LJ_AutoVocoder, trained on 24 h of adult sound data. The model trained on CTT achieved an MCD value of 1.84 for girl audio samples. For boy audio samples, the MCD value for CTT_AutoVocoder was 2.05. For LJ_AutoVocoder, the MCD values across all speakers were 1.97 for boys and 2.13 for girls.

Table 3. Mel-cepstral distortions mcd (db) average results between the transformed and original mel cepstral sequences for 12 samples produced during the experiments by both models.

Systems	Boy	Girl
CTT_AutoVocoder	2.05	**1.84**
LJSpeech_AutoVocoder	**1.97**	2.13

F0 Root Mean Square Error (F0-RMSE): In comparing log F0 values between the ground truth and synthesized waveform, we utilized F0 root mean square error (F0-RMSE) [33] as an evaluation metric. The F0-RMSE values for each model across 12 samples are reported in Table 3. The F0-RMSE is computed using the following equation:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(F0_i - \acute{F}\,0_i\right)^2}. \tag{2}$$

In the above equation, N represent the total number of frames in the speech, the term $\acute{F}\,0_i$ denotes the fundamental frequency F0 at the i^{th} frame of the generated waveform, and $F0_i$ represents the F0 value at the i^{th} frame of the ground truth. Waveform. The smaller value of F0-RMSE implies a lower prediction error. In our evaluation, we calculated RMSE values by comparing the ground truth waveform's fundamental frequency (F0) with the synthesized waveform's. F0-RMSE is used to evaluate the performance of the acoustic model, while AutoVocoder serves as the waveform generator (Table 4).

Table 4. F0 Root Mean Square Error (F0-RMSE): values for both models as an average over 12 samples for both genders.

Systems	Boy	Girl
CTT_AutoVocoder	**2.96**	3.23
LJSpeech_AutoVocoder	3.00	**3.19**

It was observed that both models exhibited convergent results in terms of F0-RMSE. Across all evaluated metrics, the AutoVocoder trained on the CTT Dataset demonstrated

exceptional performance in synthesizing highly quality sounds. It exhibited a remarkable ability to closely approximate the quality of the original sound.

4.2 Subjective Listening Test

A MUSHRA (Multiple Stimuli with Hidden Reference and Anchor) style evaluation [34] was used to compare the performance of AutoVocoder trained on both the CTT and LJSpeech Datasets. MUSHRA is more sensitive to small differences in audio quality and allows participants to directly compare multiple stimuli, making it ideal for fine-grained analysis of expressive child speech. Since expressive speech synthesis requires careful evaluation of nuances like emotion, tone, and clarity, MUSHRA provides a more detailed and precise evaluation framework. Each MUSHRA screen presented 4 stimuli to the listener for evaluation. These were CTT_AutoVocoder, LJSpeech_AutoVocoder, lower-anchor: low-pass filtered at 3.5 kHz, and ground truth. Twelve unseen samples during the training phase are used in the listening test. Twenty three listeners were instructed to evaluate and rate the stimuli based on the provided conditions. Listeners were instructed to find the ground truth within those 4 samples and assign it a score ranging from (highly not similar, not similar, intermediate similar and highly similar), whilst rating all samples. Figure 2 shows the mean naturalness scores for the models.

As illustrated in Fig. 2, the MUSHRA test results demonstrate a comparable preference among listeners for the synthesized sound samples produced by both AutoVocoders across genders Specifically, there is a similar preference for the audio samples generated by the CTT_AutoVocoder and LJspeech_AutoVocoder.

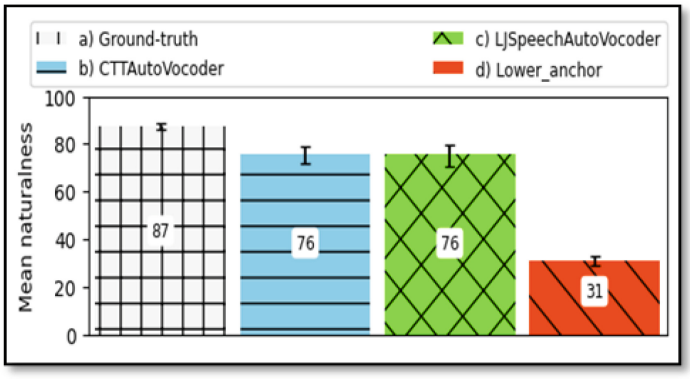


Fig. 2. The MUSHRA scores for the Mean naturalness are presented for (a) Ground truth (b) CTT_ AutoVocoder (c) LJSpeech_ AutoVocoder (d) Lower anchor, with the average results shown. A higher value indicates better overall quality.

The preferences of listeners indicated that the CTT_AutoVocoder exhibited a significant similarity to the ground truth in the MUSHRA test results. Notably, 76% of respondents perceived the CTT_AutoVocoder as very similar to the ground truth audio samples.

5 Conclusions

In this paper we introduced a new expressive CTT dataset for child speech synthesis and TTS applications. CTT includes over 1200 utterances from 25 children extracted from TEDx Talks for Kids events uploaded to YouTube. The age varies from 6 to 11 years. To verify the effectiveness of CTT in real-world situations, AutoVocoder models were trained and synthesized samples were generated. The models were trained on both LJSspeech large scale dataset and our CTT dataset. The experimental results indicate that the CTT dataset can steadily give comparable results with acoustic model trained on a large-scale dataset with a size of less than 10% of the large dataset.

Additionally, the dataset introduced here has the potential to be used for ASR systems that often struggle with child speech.

The limitations of our dataset are its small size, and the fact that it is not publicly available yet. However, the dataset is accessible for research purposes. We hope that our new dataset will be adopted, alongside other child datasets as a benchmark in the speech processing research community to train models for child speech. In future work, we can further augment the dataset through augmentation techniques, employing it in the development of fully expressive child text to speech synthesis and enhance the duration by increasing the number of hours. We will focus on addressing age-based differences in TTS performance, as speech synthesis models.

Acknowledgement. This paper is supported by the European Union's HORIZON Research and Innovation Programme under grant agreement No 101120657, project ENFIELD (European Lighthouse to Manifest Trustworthy and Green AI) and by the Ministry of Innovation and Culture and the National Research, Development and Innovation Office of Hungary within the framework of the National Laboratory of Artificial Intelligence.

Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union and the granting authorities. Neither the European Union nor the granting authorities can be held responsible for them.

References

1. Shaheen, Z., Sadekova, T., Matveeva, Y., Shirshova, A., Kudinov, M.: Exploiting emotion information in speaker embeddings for expressive text-to-speech. In: INTERSPEECH, pp. 2038–2042 (2023)
2. Zhao, W., Yang, Z.: An emotion speech synthesis method based on vits. Appl. Sci. **13**(4), 2225 (2023)
3. Meng, Y., et al.: CALM: contrastive cross-modal speaking style modeling for expressive text-to-speech synthesis. arXiv preprint arXiv:2308.16021 (2023)
4. Perrotin, O., Stephenson, B., Gerber, S., Bailly, G.: The blizzard challenge 2023. In: 18th Blizzard Challenge Workshop, ISCA, pp. 1–27 (2023)
5. Xu, Z., et al.: MuLanTTS the microsoft speech synthesis system for blizzard challenge 2023. arXiv preprint arXiv:2309.02743, (2023)
6. Hagen, A., Pellom, B., Cole, R.: Highly accurate children's speech recognition for interactive reading tutors using subword units. Speech Commun. **49**(12), 861–873 (2007)

7. Terblanche, C., Harty, M., Pascoe, M., Tucker, B.V.: A situational analysis of current speech-synthesis systems for child voices: a scoping review of qualitative and quantitative evidence. Appl. Sci. **12**(11), 5623 (2022)
8. Ward, W.: My science tutor and the MyST corpus (2019). https://www.researchgate.net/publication/331210819
9. Yeung, G., Fan, R., Alwan, A.: Fundamental frequency feature warping for frequency normalization and data augmentation in child automatic speech recognition. Speech Commun. **135**, 1–10 (2021)
10. Eskenazi, M., Mostow, J., Graff, D.: The CMU kids corpus. In: Linguistic Data Consortium, vol. 11 (1997)
11. Shobaki, K., Hosom, J.-P., Cole, R.: The OGI kids' speech corpus and recognizers. In: Proceedings of ICSLP, Citeseer, pp. 564–567 (2000)
12. Kazemzadeh, A., et al.: TBALL data collection: the making of a young children's speech corpus. In: Proceedings of the INTERSPEECH, pp. 1581–1584 (2005)
13. Demuth, K., Culbertson, J., Alter, J.: Word-minimality, epenthesis and coda licensing in the early acquisition of English. Lang. Speech **49**(2), 137–173 (2006)
14. Lo, T.-H., Chao, F.-A., Weng, S.-Y., Chen, B.: The NTNU system at the interspeech 2020 non-native children's speech ASR challenge. arXiv preprint arXiv:2005.08433 (2020)
15. TEDx Talks. https://www.youtube.com/@TEDx
16. Ito, K., Johnson, L.: The lj speech dataset (2017). https://keithito.com/LJ-Speech-Dataset
17. Matassoni, M., Falavigna, D., Giuliani, D.: DNN adaptation for recognition of children speech through automatic utterance selection. In: 2016 IEEE Spoken Language Technology Workshop (SLT), pp. 644–651. IEEE (2016)
18. Shivakumar, P.G., Georgiou, P.: Transfer learning from adult to children for speech recognition: evaluation, analysis and recommendations. Comput. Speech Lang. **63**, 101077 (2020)
19. Matassoni, M., Gretter, R., Falavigna, D., Giuliani, D.: Non-native children speech recognition through transfer learning. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6229–6233. IEEE (2018)
20. Tong, R., Wang, L., Ma, B.: Transfer learning for children's speech recognition. In: 2017 International Conference on Asian Language Processing (IALP), pp. 36–39. IEEE (2017)
21. Gerosa, M., Giuliani, D., Brugnara, F.: Acoustic variability and automatic recognition of children's speech. Speech Commun. **49**(10–11), 847–860 (2007)
22. Fainberg, J., Bell, P., Lincoln, M., Renals, S.: Improving children's speech recognition through out-of-domain data augmentation. Interspeech **2016**, 1598–1602 (2016)
23. Hasija, T., Kadyan, V., Guleria, K.: Out domain data augmentation on Punjabi children speech recognition using Tacotron. J. Phys. Conf. Ser. **1950**, 012044 (2021)
24. Serizel, R., Giuliani, D.: Deep neural network adaptation for children's and adults' speech recognition. In: Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 and of the Fourth International Workshop EVALITA 2014 9-11 December 2014, Pisa, pp. 344–348 (2014)
25. Alm, C.O., Sproat, R.: Perceptions of emotions in expressive storytelling. In: 9th European Conference on Speech Communication and Technology, pp. 533–536 (2005). https://doi.org/10.21437/interspeech.2005-334
26. Harikrishna, D.M., Gurunath Reddy, M., Rao, K.S.: Multi-stage children story speech synthesis for Hindi. In: 2015 Eighth International Conference on Contemporary Computing (IC3), pp. 220–224. IEEE (2015). https://doi.org/10.1109/IC3.2015.7346682
27. Lakomkin, E., Magg, S., Weber, C., Wermter, S.: KT-speech-crawler: automatic dataset construction for speech recognition from YouTube videos. arXiv preprint arXiv:1903.00216 (2019)

28. Li, X., et al.: IEEE automatic speech recognition and understanding workshop (ASRU). IEEE **2023**, 1–8 (2023)
29. Boersma, P.: Praat: doing phonetics by computer (2007). http://www.praat.org/
30. Webber, J.J., Valentini-Botinhao, C., Williams, E., Henter, G.E., King, S.: Autovocoder: fast waveform generation from a learned speech representation using differentiable digital signal processing. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5. IEEE (2023). https://doi.org/10.1109/ICASSP49357.2023.10095729
31. Alwaisi, S., Al-Radhi, M.S., Németh, G.: Automated child voice generation: methodology and implementation. In: 2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), pp. 48–53. IEEE (2023)
32. Zhang, M., Zhou, Y., Zhao, L., Li, H.: Transfer learning from speech synthesis to voice conversion with non-parallel training data. IEEE/ACM Trans. Audio Speech Lang. Process **29**, 1290–1302 (2021). https://doi.org/10.1109/TASLP.2021.3066047
33. Luo, Z., Chen, J., Takiguchi, T., Ariki, Y.: Emotional voice conversion with adaptive scales F0 based on wavelet transform using limited amount of emotional data. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp. 3399–3403. International Speech Communication Association (2017). https://doi.org/10.21437/Interspeech.2017-984
34. Recommendation, I.: 1534–1, 'Method for the subjective assessment of intermediate sound quality (MUSHRA)'. In: International Telecommunications Union, Geneva, Switzerland, vol. 2 (2001)

Multidimensional Rhythm: Comparing Rhythmic Properties of Australian and New Zealand Monologues

Anna Borzykh(✉) and Tatiana Shevchenko

Moscow State Linguistic University, 38 Ostozhenka Street, Moscow 119034, Russian Federation
anna.a.borzykh@mail.ru

Abstract. The study is concerned with comparison of rhythmic features in monologues of Australian (AusE) and New Zealand (NZE) speakers. Methodology is designed to capture the advantages of both the traditional approach and the 'new paradigm' metrics to account for the two national varieties of English. The traditional for Russian linguistics description of rhythm as a hierarchy of linguistic units (ip, foot, syllable) provided data on common rhythm structure with dialect-specific differences in syllable durations, which suggested the Australian speech habits of prolonging unstressed syllables. By applying the metrics collected in Correlatore [15], we found that unmonitored monologues' rhythm could be categorized differently from reading reported in previous research. AusE speech demonstrated features of 'controlling' syllable-based rhythm, in contrast with the 'compensating' accent-based rhythm of NZE speakers. The metrics proved to be dialect-, style- and tempo-sensitive, and suitable for comparison of varieties of the same language.

Keywords: Australian English · New Zealand English · Rhythm Metrics · Accent-Based Rhythm · Syllable-Based Rhythm

1 Introduction

The aim of the present study is to compare rhythmic features in two lesser-known varieties of English, namely Australian English (AusE) and New Zealand English (NZE). These varieties are known to be very similar, and research confirms that speakers of other varieties, such as British English and American English may not be able to distinguish speakers of the two varieties in question [14]. Nevertheless, special research in prosody revealed that there are differences between them, though nuanced rather than categorical, which might be showing different trends in the two varieties [4].

It is generally accepted that Australian intonation patterns are marked by the monotony of pitch variation, by a narrow pitch range and a high frequency of level tones as well as levelled out sliding patterns in the preterminal parts of ip [21]. Widely commented is the use of rising tones in narratives, the so-called High Rising Tone (HRT) in statements, which is typical both of AusE and NZE [13]. It is indisputable that Australian English has features of British south-eastern low class pronunciation habits, which

A. Karpov and V. Delić (Eds.): SPECOM 2024, LNAI 15299, pp. 241–250, 2025.
https://doi.org/10.1007/978-3-031-77961-9_18

is confirmed by the retrospect data gathered about Cockney and its prosody. In particular, the accent-based rhythm of today's Cockney English in the South-East of England was brought and preserved in the territory of Australia, while in NZE it is language contact with mora-based rhythm of Maori which is expected to affect the pronunciation of Pakeha (the population of European origin).

New developments of the two varieties are reported being concerned with Greek and Italian speakers in Australia, as opposed to Maori speakers in New Zealand [14]. These two groups appear to be reference groups for the innovations which might influence the development of prosody – and pronunciation in general – in either country. AusE is noted for new tendencies due to Greek and Italian residents' influence, compared with NZE having the powerful impact of Maori, one of the indigenous peoples in New Zealand [14]. The socio-historical backgrounds of the two varieties are different, especially since the times of migration and settlements were different. Nevertheless, AusE might be the one which had more impact on NZE than the Maori language contact.

AusE and NZE exhibit a number of nationally specific temporal characteristics. As evidenced by previous research data, for instance, NZE has a faster tempo than AusE [16]. As a result, accentuation which is due to prominence of stressed syllables compared with unstressed syllables might be achieved by different acoustic means in the two varieties. It was also found that the choice of the acoustic correlate of accented syllables' prominence was gender-specific. Australian men proved to rely to a greater extent on intensity, while Australian women rely more on pitch. Common for both genders was contrastive duration of stressed and unstressed syllables, which also served as a feature of prominence. Given the importance of duration, we set ourselves the task to verify the point of timing in accentual prominence in AusE and NZE.

The previous findings of a specifically rhythmic experiment based on a large group of speakers [5] confirmed the accent-based rhythm of Pakeha, i.e. the white population or people of European origin, compared with the Maori, which is closer to syllable-based [19]. To be exact, the Maori language is mora-based, but the smaller and more detailed quantifying of syllables might be very similar for the mora-based and syllable-based rhythm. Further research aimed at measuring Australian rhythm included the combination of three metrics and was backed up by Szakay's data [5]. The results demonstrated that AusE is undoubtedly accent-based, with Pakeha values being very close to it, while Maori is definitely syllable-based. The data on NZE borrowed from Szakay [19] appeared to be fruitful and productive, as it can show the comparison between the two varieties, suggesting that further developments of New Zealand rhythm might be in the line of becoming more syllable-based. This tendency is mostly typical of speakers who are greatly involved in the Maori culture, which probably was not the case of the Pakeha speakers, whose values were close to AusE accent-based data.

Before covering the methodology details, one more idea should be made clear. The whole point of the research undertaken in the two varieties is to show that, like other authors [10, 11], we do not approve of oversimplification in categorizing the two varieties of the same language as either accent-based or syllable-based. We predict that there may be certain tendencies for change in the direction of either category which become transparent depending on the parameters of the research. It is crucial to understand that we deal with certain grades of one and the same quality placed in a continuum that could

vary according to the context of the situation, i.e. display its adaptability to style and individual speech habits.

2 Methodology

The Material. The novelty of the present study is that the two varieties were to be compared on the basis of homogeneous material of natural unmonitored speech of short narratives. We collected the data through internet: 12 volunteers from Australia (6) and New Zealand (6) agreed to record their 2 min monologues about themselves or an episode in their language learning. The total time is 24 min produced by 7 men and 5 women, young people, aged 20–28. Personal names were elicited or changed and geographical place names were clarified.

The Traditional Method. The aim of the present research is to cover as many features of rhythm as possible catching the advantages of the two approaches. One is traditional for Russian linguistics; it defines rhythm as a periodicity of linguistic units, such as the syllable, the foot (also known as accent group / rhythmic group / stress group), and the ip (which stands for a pause-to-pause period of phonation, also known as 'syntagma', or 'intonation phrase' in the present-day terminology). The advantage of the traditional linguistic division into ips, feet and syllables consists in the fact that we get a hierarchy of linguistic units. The results confirmed that each little element (such as the syllable, for instance) is nested within the foot, and the foot is nested within the ip. Therefore, there is a certain rhythmical arrangement of linguistic units having a syntactically determined structure. It was also stated that rhythm units were biologically and cognitively determined, with ip duration being constrained by breathing, syllable duration being similar to heartbeat and the supra-phrasal unit bearing on one topic. For each member in the hierarchy a certain range of timing was established: for ip it was 1–2 s, for the foot it ranged from 400 ms to 1000 ms, and the average syllable duration was set at 200 ms [1].

Measurements. In the current research measurements were performed manually in Praat [3] and included the duration of ips, feet and syllables. Additional measurements were concerned with accented and unaccented syllables' duration. The division into three degrees of accentual prominence (primary, secondary and unstressed) was based on two national dictionaries, the Australian and the New Zealand ones [6, 9], supported by audio-visual observations made by two experienced labelers with 81.5% agreement.

Average syllable duration (ASD) is taken to represent the rate of articulation which proved to be significant for comparing the two dialects.

The New Paradigm Method. Segmentation and *annotation* of the speech signal into vocalic and consonantal intervals was done manually in Praat [3] as suggested by previous research in [15], using CV annotation system in which consonants are marked as C, vowels and approximants as V, with # for a pause. The aligned and annotated speech samples were saved in TextGrid format which was then fed into the Correlatore program [15]. The new paradigm metrics, which are collected in Correlatore, include the deltas [17], the PVIs [12], the Varcos [8], and last but not least, the CCI method [2]. The latter divides speech samples into the so-called 'compensating' and 'controlling' types of rhythm. The 'compensating' type means compensating the length of the sounds within

a syllable, as well as compensating the length of the adjacent feet, or adjacent syllables in a foot. It is common knowledge that the more syllables are included into the foot, the more compressed and therefore shortened their length or duration will be. Compensating features are included in the accent-based rhythm. 'Controlling' means having more or less equal length of syllables, which is more obvious in the case of syllable-based rhythm of most other languages. The authors argued that their method might suit the comparison of dialects of the same language; they also predicted that the outcome may be style- and tempo-dependent, as well as reflecting individual speakers' speech habits [2].

Factors which Affect Rhythm. Already within the framework of the traditional approach it was found that the interplay of different speech units' timing could have its impact on the rhythmic patterns of one language. The first most noticeable factor was discourse type, or style, which was demonstrated by prose reading, verse and spontaneous speech [1]. Another factor was time in language acquisition and human speech development across the lifespan [18]. D. Crystal, for instance, titled his paper "Documenting rhythmical change" and argued that one particular individual could develop and vary one's timing in different situations and at different periods of life. The author also gave examples of socially motivated change in English rhythm, like syllable timing in Airspeak spoken between pilots and ground service or rap performance popular today with the young [7]. In the present corpus we collected speech samples of one particular discourse type and limited the age of speakers to the young people aged 20–30.

Combining the Two Approaches. In our methodology we follow the wholistic view of the language structure aiming at collecting maximum of relevant information on rhythm, starting from the major composition of recurrent speech units and finishing with fine-grained comparisons of syllables, their vocalic and consonantal constituents, and their relative durations. In other words, we hypothesize that by combining the traditional and the latest approaches we can achieve a fair view of what constitutes the identity of AusE speakers compared to NZE speakers in rhythm.

Statistical Analysis. Finally, to prove the reliability of the obtained data as well as to draw reasonable conclusions, statistical analysis was used. The performed tests in Jamovi include the Shapiro-Wilk's normality test, one-way ANOVA (non-parametric), Welch's one-way ANOVA, and the Games-Howell post-hoc test [20].

3 Results

3.1 The Results of the Traditional Approach

Our first observation is concerned with the total amount of ips, feet and syllables, as well as their relative durations. As predicted, the hierarchy of the rhythmic units is preserved in both Australian English and New Zealand English. The linguistic units are hierarchically structured at phrasal, clausal, word and syllable levels: an ip consists approximately of two feet (for AusE in the range of 2.0–2.6, for NZE in the range of 2.0–2.5); a foot likewise consists of two syllables (for AusE in the range of 2.0–2.6, for NZE in the range of 2.0–2.4). That symmetrical arrangement (see Fig. 1) of similar

proportions does not show, however, the difference between the two national variants in articulation rate, which could be relevant for rhythm perception.

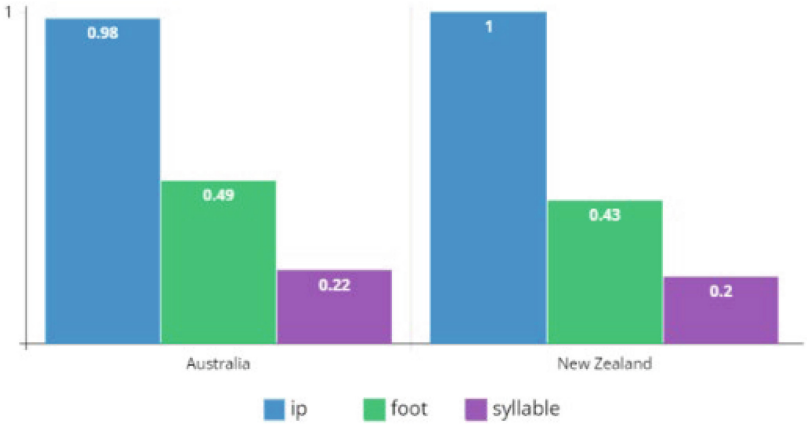


Fig. 1. Median duration of ips, feet and syllables.

By applying the Shapiro-Wilk's normality test and one-way ANOVA (non-parametric) we found that average syllable duration (ASD) data as one of the basic features of tempo (together with pause duration) gave evidence of NZE average syllable duration being shorter, which is a sign of faster tempo (see Fig. 2).

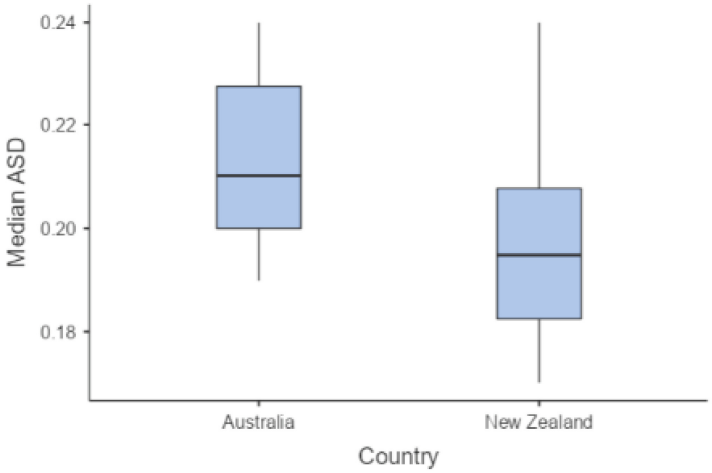


Fig. 2. Median ASD.

Furthermore, a number of polysyllabic words were selected from the sounding material both by Australian speakers (26 words) and New Zealand speakers (13 words) in order to calculate the duration of accented and unaccented syllables. Three degrees of

word stress were differentiated, namely primary stress, secondary stress and unstressed syllables. Average and median values were computed and compared statistically [20] to examine the contrast in their duration (see Table 1). The contrast between syllables bearing primary stress and those bearing no stress was of the main importance.

Table 1. Duration of accented and unaccented syllables in polysyllabic words.

	primary stress (in sec)	secondary stress (in sec)	unstressed (in sec)
Australia	0.204 average 0.212 median	0.176 average 0.176 median	0.160 average 0.150 median
New Zealand	0.207 average 0.189 median	0.150 average 0.145 median	0.150 average 0.124 median

Syllables with primary stress, secondary stress and unstressed ones exhibit contrasting average durations both in Australia ($F = 4.04$, $p = 0.036$) and in New Zealand ($F = 4.26$, $p = 0.026$) according to Welch's one-way ANOVA. However, when it comes to comparing the duration of primary stressed syllables to unstressed syllables, a less sharp contrast is observed in AusE (44 ms) as compared to NZE (57 ms), which signals AusE being closer to syllable-based rhythm while NZE to accent-based rhythm. The significance of the obtained results was tested by means of the Games-Howell post-hoc test, which confirmed the difference between primary stressed syllables and unstressed syllables in AusE ($p = 0.014$), as well as in NZE ($p = 0.056$, marginal). As for the difference between syllables bearing primary stress and secondary stress, a sharper contrast is observed in NZE (57 ms, $p = 0.027$), with a less sharp one in AusE (28 ms, $p = 0.520$, marginal). Finally, the comparison of syllables bearing secondary stress and unstressed ones yielded no significant differences either in AusE and NZE.

We can, therefore, state with confidence that application of the first traditional approach yielded relevant linguistic information: a) about the hierarchy of linguistic units whose durations of dual nature appear to be rhythmically structured, which is a feature shared by both AusE and NZE; b) in AusE average syllable duration is greater, which is a sign of slower articulation rate, while a lower articulation rate level in NZE testifies to a relatively faster tempo; c) rhythmically valid are comparisons in contrast between accented and unaccented syllables; our preliminary results suggest that AusE does not make that contrast impressive enough to produce the effect of prominence due to accent timing.

3.2 The Results of the Modern Approach

The data obtained by applying the metrics are presented both numerically and visually by means of scattergrams. Figure 3 reflects the relatively greater saturation with vowels and greater variability of vocalic intervals in the speech signal in AusE, as compared with NZE.

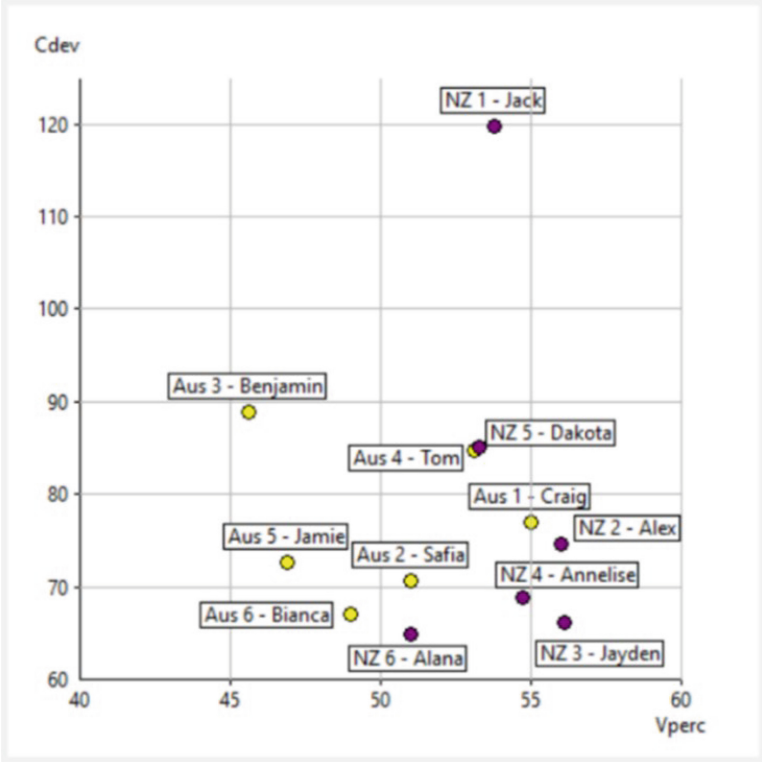


Fig. 3. %V – ΔC.

Categorical differences between the two national varieties are best revealed by the CCI metric: most of the AusE speakers' scores cluster above the bisecting line, whereas most NZE speakers' scores are located below the bisecting line. The expected interpretation of the results is that we received evidence that AusE tends to be syllable-based in narratives, while NZE remains in the accent-based category. Despite faster articulation rate NZE speakers produce higher vocalic fluctuations, probably at the expense of unstressed vowels' reduction (see Fig. 4).

In the group of NZE speakers, however, we can observe two outliers whose speech habits are different from the general trend. We can account for those cases by assuming that the individual variability is associated with greater involvement in Maori culture and communication, as was suggested by previous research [19].

4 Conclusions and Discussion

The present study demonstrated the complex nature of speech rhythm which could be measured along different lines by selecting a number of dimensions: the major division into syntactically relevant units at syllable, word and clause levels, on the one hand, and phonologically relevant vowels and consonants in the syllable, on the other. Both the hierarchy of linguistic units and the relative proportions of vowels and consonants within and between the syllables prove to be rule-based, specific for rhythmic structures of particular varieties in the same language. By looking at vowels and consonants' variability we can observe the mechanism of constructing rhythmic units of a higher order.

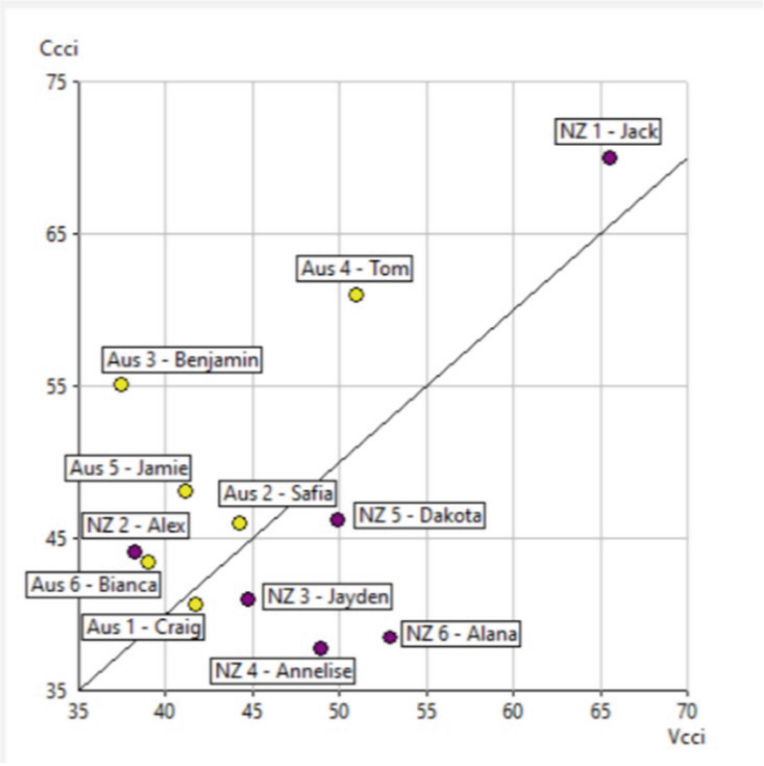


Fig. 4. vCCI – cCCI.

AusE and NZE are two national varieties of the same language whose populations had their specific socio-demographic origins and language contacts which affected their speech habits, including rhythm. Nevertheless, in reading their common feature is the accent-based English rhythm, as evidenced by previous research data [5, 19], while in monologues spoken in a relaxed manner Australian speakers tend to produce higher scores in average syllable duration (slower tempo), less contrast between stressed and

unstressed syllables, and, as a result, display syllable-based rhythm. NZE speakers, unless they are involved in the Maori culture, tend to preserve accent-based rhythm in monologues despite their faster tempo.

The multidimensional approach showed the mechanism of inter-level dependence, the bond between the segmental and the suprasegmental rhythm components.

The present research demonstrated the influence of discourse setting and the style of speech on rhythm measurements in AusE and NZE. Other factors which affect rhythm might be age and gender, mentioned in the previous research data. The limitations of the present research consist in a relatively small number of speakers. Further research might provide data on a larger corpus of respondents from both countries engaged both in reading and speaking.

References

1. Antipova, A.M.: Ritmicheskaya sistema angliiskoi rechi (Rhythmic System of English Speech). Moscow. Vysshaya shkola. (in Russian) (1984)
2. Bertinetto, P.M., Bertini, C.: On modeling the rhythm of natural languages. In: Barbosa, P.A., Madureira, S., Reis, C. (eds.) Proceedings of Speech Prosody 2008, Campinas (Brazil), 6–9 May 2008, pp. 427–430 (2008)
3. Boersma, P., Weenink, D.: Praat: doing phonetics by computer [Computer program] (2024). http://www.praat.org/
4. Borzykh, A.A.: Comparative analysis of Australian English and New Zealand English national standards' prosodic features. Vestnik of Moscow State Linguistic University. Humanities. **8**(863), 9–14. https://doi.org/10.52070/2542-2197_2022_8_863_9 (2022)
5. Buraya, E.A.: Prosodic rhythm in Australian English (Gender differentiation). Teoretich-eskaya i prikladnaya lingvistika [Theoretical and Applied Linguistics], **7** (4), 5 (2021). https://doi.org/10.22250/24107190_2021_7_4_5_15
6. Butler, S. (ed.): Macquarie Concise Dictionary (6th edition). – Sydney: Macquarie Dictionary Publishers (2013)
7. Crystal, D.: Documenting rhythmical change. In: Windsor Lewis, J. (ed.), Studies in general and English phonetics (London: Routledge), pp. 174–9 (1994)
8. Dellwo, V., Wagner, P.: Relations between language rhythm and speech rate. In: 15th International Congress of Phonetic Sciences (ICPhS) (2003)
9. Deverson, T., Kennedy, G. (eds.): The New Zealand Oxford Dictionary. Oxford University Press, Oxford (2005)
10. Fuchs, R.: Analysing the speech rhythm of New Englishes: a guide to researchers and a case study on Pakistani, Philippine, Nigerian and British English. In: Wilson, G., Westphal, M. (eds.) New Englishes, New Methods, pp. 132–155. Benjamins, Amsterdam (2023)
11. Gibbon, D., Gut, U.: Measuring speech rhythm. In: Interspeech (2001)
12. Grabe, E., Low, T.L.: Durational variability in speech and the rhythm class hypothesis. In: Gussenhoven, C., Waner, N. (Eds.), Papers in Laboratory Phonology 7, pp. 515–546. Cambridge University Press (2002)
13. Jun, S.A. (ed.): Prosodic Typology. Oxford University Press, The Phonology of Intonation and Phrasing. Oxford (2005)
14. Kiesling, S.F.: English in Australia and New Zealand. The Handbook of World Englishes, pp. 70–86 (2020). https://doi.org/10.1002/9781119147282.ch5

15. Mairano, P., Romano, A.: Un confronto tra diverse metriche ritmiche usando Correlatore. In: Schmid, S., Schwarzenbach, M. & Studer, D. (eds.) La dimensione temporale del parlato, (Proc. of the V National AISV Congress, University of Zurich, Collegiengebaude, 4–6 February 2009), Torriana (RN): EDK, pp. 79–100 (2010)
16. Nokes, J., Hay, J.: Acoustic correlates of rhythm in New Zealand English: a diachronic study. Lang. Var. Change **24**, 1 – 31 (2012)
17. Ramus, F., Nespor, M., Mehler, J.: Correlates of linguistic rhythm in the speech signal. Cognition **73**(3), 265–292 (1999)
18. Shevchenko T., Sokoreva T.: Starting a conversation: indexical rhythmical features across age and gender (a corpus study). In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) Text, Speech, and Dialogue. 19th International Conference, TSD 2016 Brno, Czech Republic, September 12–16, 2016 Proceedings. LNAI 9924 Lecture Notes in Artificial Intelligence, pp. 495–505. Springer, Cham (2016)
19. Szakay, A.: Rhythm and pitch as markers of ethnicity in New Zealand English. In: Warren, P., Watson, C.I. (eds.), Proceedings of the 11th Australian International Conference on Speech Science and Technology, pp. 421 – 426. University of Auckland, New Zealand (2006)
20. The jamovi project. jamovi. (Version 2.3) [Computer Software] (2022). https://www.jamovi.org
21. Typology of Variants of the English Phonological System. In: Shevchenko, T.I. (ed.). Tula University Press (in Russian) (2012)

Influence of Linguistic and Sociolinguistic Factors on Speech Rate Perception

Anastasia Ananeva(✉) and Uliana Kochetkova

Saint Petersbourg State Univerity, Saint Petersbourg, Russia
st075648@student.spbu.ru, u.kochetkova@spbu.ru

Abstract. This study aims to investigate the perception of speech rate in Russian. The goal is to examine the impact of pauses duration, articulation rate, as well as speaker's and listener's gender on speech rate perception. The relevance of this study lies in its potential to improve speech synthesis systems by understanding how speech rate perception varies, particularly between genders. The novelty of the research is in its comprehensive analysis of the impact of gender, articulation speed, and pauses on speech rate perception based on Russian language material. The material of the current study consists of recordings of 2 texts read by 55 native Russian speakers. The series of auditory perceptual experiments contained both natural and modified stimuli. The results of the experiments revealed the impact of speaker's gender on speech rate perception, but no impact of listener's gender. We also found the effect of pauses duration on overall speech rate perception. The results of the study showed that we can describe the speech rate perception through its intrinsic characteristics, such as pausing and articulation rate. We also proposed a model of speech rate perception in Russian language. Although this study did not focused on the effect that speech style and other factors may produce, the data obtained allow identifying important patterns in listener's behavior, that can be applied in natural language processing, as well as in artificial intelligence systems.

Keywords: Speech rate · Articulation rate · Perception · Gender

1 Introduction

The perception of speech rate is a complex process, necessitating the consideration of human perceptual characteristics and an analysis of both linguistic and social dimensions. Speech rate is composed of articulation rate and pauses and is influenced by extra-linguistic, paralinguistic and language-relevant factors [28], which create its complex structure.

Summarizing findings from various studies, it becomes evident that perception results from the interplay of numerous factors affecting an individual. The perception of time and speed, in particular, is unique among sensory experiences because there is no specific organ responsible for it [25]. Major theories of time

A. Karpov and V. Delić (Eds.): SPECOM 2024, LNAI 15299, pp. 251–264, 2025.
https://doi.org/10.1007/978-3-031-77961-9_19

perception include biological [2] and cognitive [21] perspectives. Understanding perception necessitates familiarity with basic psychophysical laws, such as the Weber-Fechner law or Stevens' law [1,20,30]. The perception of speech rate can be aligned with the latter, requiring the determination of an exponent that could describe speech rate perception.

Researchers have found that the perception of speech rate is influenced by various factors, including gender, age, speech and hearing disorders, and listener's own speech rate, among others [4–6,9,15,24]. The discrimination threshold for speech rate is about 5% [8,23], indicating high sensitivity to changes in rate. Produced speech rate is a factor that directly influences perception: those who speak slowly tend to overestimate the speech rate of others, particularly among those who speak relatively slower [24].

Additionally, men are perceived as speaking faster than women [9]. There is also a correlation between individual and perceived speech rates with the listener's gender. Men who speak faster tend to underestimate the speech rate of women, and vice versa. Some authors suggest that speech rate perception is more closely related to voice pitch than to the speaker's gender [5].

When discussing speech rate, it is important to introduce a number of concepts. First and foremost, it is necessary to distinguish between articulation rate and speaking rate or overall speaking rate. In the first case, the rate is measured without considering pauses; that is, only the time spent directly on articulation is counted. Speaking rate, on the other hand, refers to the total time required to produce all elements of speech, including pauses.Therefore, pausing and articulation rate [10–12,14,18] are the main characteristics of speech rate that largely determine its perception. Therefore, the perceived speech rate can be described as a function of these two elements. Lane and Grosjean [12] came up with the following formula after conducting several experiments based on English material:

$$E' = A^{-2} + 6P^{-0.2} - 10 \tag{1}$$

The coefficient corresponding to articulation rate is –2, and –0.2 for pausing, which demonstrates the impact of the two components. These data enable the construction of the model which describes speech rate perception with an accuracy of up to 0.96. The authors suggest that the model is applicable for almost any language.

The present study aims to find differences in the perception of speech rate, which depend on the gender of speakers and describe the models of speech rate perception that is suitable for Russian language.

2 Gender Influence on the Perception of Speaking Rate and Articulation Rate

2.1 Methodology and Materials

Four pairs of speakers (male and female) with approximately the same overall speech rate were selected from 53 recordings of a text read by native Russian

speakers from the CoRuss corpus for the first experiment. CoRuss is a corpus of Russian spontaneous speech which consists of dialogues between, monologues and reading of a short phonetically balanced text [13]. As the difference in tempo between pairs is less than 0.5%, which is less than the threshold of distinction [23], there are no objective indicators that speakers speak at different speeds. Two phrases identical for each pair of speakers were selected from the recordings. Recordings in pairs were brought to the same duration by increasing or decreasing the length of pauses at the beginning or the end. Participants were asked to listen to eight pairs of recordings (two of each speaker's recordings), ranging in duration from 7 to 13 s, and to determine which of them sounded faster. The survey was conducted using the SoSci Survey platform. A total of 37 auditors participated in the experiment: 25 females and 12 males aged from 16 to 52.

For the second perceptual experiment we selected recordings of the speakers with the similar articulation rate. For each pair of speakers, phrases of 2 to 5 s in length were selected. In most cases, they consisted of 1 syntagm, so there were no pauses in them. To minimize the pause influence (if there are any) the recordings were modified: the duration of all pauses within the selected fragments was brought to the same length by adding or removing part of the pause. All audio recordings were equalized to the same duration. Twenty-six people took part in the survey: 12 men and 14 women aged from 18 to 25.

2.2 Results

The results show no significant influence of speaker's or listener's gender on the perception of overall speaking rate. However, a deeper analyses of the results of each question showed that men tend to have difficulties rating the recordings presented if one of them contains more hesitation elements. Only 8% of women found it difficult to give an answer for question 7, which includes a great amount of hesitation pauses, while almost half of men did. This distribution can be called unusual, as in all other questions the percentage of men or women who found it difficult to give an answer does not differ much (Table 1).

Table 1. The results of statistical analysis.

Criteria	Method	p-value
Participant gender influence on perception of speaking rate	χ^2	p = 0,31
Pause lengths influence	χ^2	p < 0,001
Speaker gender influence on perception of speaking rate	t-test	p = 0,32
Participant gender influence on perception of articulation rate	χ^2	p = 0,28
Speaker gender influence on perception of articulation rate	t-test	p = 0,009

It is likely that the fact that the pauses were filled with extraneous sounds rather than emptiness made the male participants hesitate in their response.

Such an observation stands in contradiction with the claim that men are better at discriminating sound in noise [17,19]. Obviously, then, speech noise is of a different nature and, accordingly, is processed by different mechanisms.

It should be noted that some participants reported that it was often difficult for them to estimate the rate because it is unclear whether to focus on the length of pauses, the number of pauses, or the speed of articulation. Nevertheless, it can be assumed that the majority relied on the length of pauses: the bigger the difference in relative pause length between the recordings presented, the more participants choose the same option as the fastest one. In order to compare the phrases with each other, the ratio of pause time and articulation time per number of syllables was calculated, since the phrases differ in duration. Chi-square method showed a strong connection between the answers of participants and relative time spent on pauses. Thus, despite the fact that articulation tempo is considered a significant factor in speaking rate estimation, it appears that the role of this factor was markedly reduced when the two recordings are compared directly.It is much easier for a person to analyze the duration of a pause.

Participant's gender influence on the perception of articulation rate was not found either. However, the gender of the speaker has an impact on how a person perceives rate which corresponds with some previous studies [4]. Figure 1a shows that in a greater number of cases it is the male voice that is evaluated as faster. Comparison of responses of male and female groups of auditors (Fig. 1b, 1c) confirms both conclusions - the gender of the auditor has no effect on the choice (the distribution of responses in the male and female groups is approximately the same), and men are perceived to be faster.

3 A Multilinear Model of Speech Rate Perception

3.1 Methodology and Materials

To determine which aspect of speech rate (articulation rate or pauses) has the most significant impact, an experiment was designed. Two native Russian speakers, one male and one female, both aged 22 were asked to read a text titled *Pop Fan* from Grosjean and Lane's works [11,12,15] translated into Russian:

> "Что касается меня я вполне обычный пятнадцатилетний подросток не совсем сумасшедший и ничем не лучше других я слушаю радио Люксембург мои волосы модно уложены и я ношу свитера с воротником поло но я не считаю себя большим поклонником поп-музыки"

The text contains 39 words and 89 syllables. The speakers were instructed to read the text at normal, slow, and fast rates. Initially, punctuation was omitted to avoid influencing their natural pause patterns. After that, they read it again with intentional pauses inserted at various points. The total recording time was 9 min for the male speaker and 9 min 30 s for the female speaker. The recordings were segmented into pauses and speech intervals using WaveAssistant software, and a Python script was written to analyze the segmented data.

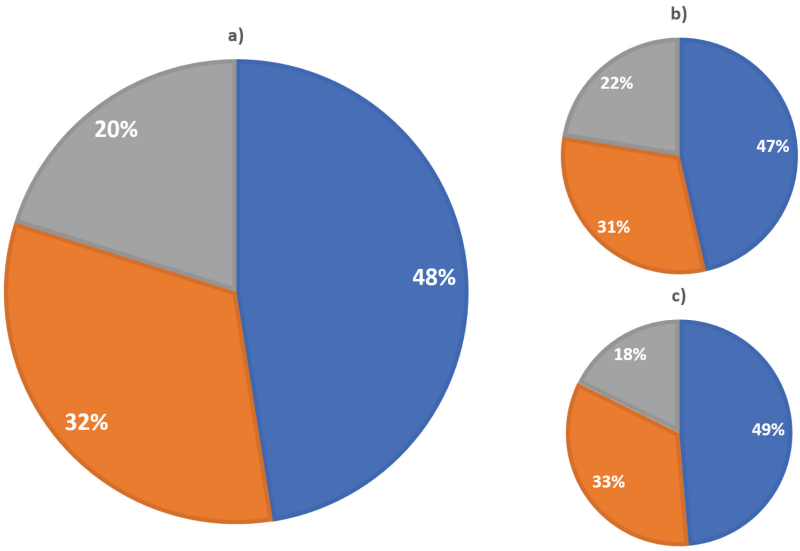


Fig. 1. Ratio of answers of a) all participants, b) women, c) men, blue indicates the number of answers "male voice is faster", orange - "female voice is faster", gray - "difficult to answer". (Color figure online)

We selected samples reflecting slow, medium, and fast articulation rates for each speaker from the recordings. Each sample was then modified to include 3, 7, and 10 pauses by either adding or removing pauses as needed. In some instances, we transferred pauses along with the final syllable of the preceding word from a similar articulation rate recording to maintain natural intonation pattern. All modified recordings underwent auditory analysis to ensure natural speech flow.

A Python program was developed to adjust the pause duration automatically. The average pause durations were set to 0.26 s, 0.49 s, and 0.72 s for the female speaker, and 0.24 s, 0.48 s, and 0.71 s for the male speaker, corresponding to the speaker's average pause duration ± 2 standard deviations.

The longest pause was Pause 4, while Pauses 2, 8, and 10 were the shortest, reflecting the text's semantic structure. Each pause was represented by a segment of zero-value signal replacing the original pause, with a minimum duration of 150 ms [29].

A total of 27 audio recordings for each of 2 speakers were created for a perceptual experiment. Due to the additional pauses, the rate of some recordings was altered. These differences did not exceed 3%, ensuring comparable tempo within each group. The durations of all recordings were standardized to eliminate reliance on absolute time metrics.

Participants were instructed to rate the recordings presented using Stevens' magnitude estimation method [26]. The instructions were following:

> "The first recording has a standard speech rate and the second recording has a changed rate. The standard one is rated as 10, and your task is to rate the changed one. In other words, the question is: if the first recording is rated as 10, how would you rate the second one? Use whatever values seem appropriate to you - fractions, decimals, or integers. For example, if the modified seems 7 times faster than the normal, put 70. If it sounds five times slower, put 2; if 20 times slower, put 0.5, etc.
> Try not to think about the sequence; your task is to assign an appropriate score to each recording, regardless of how you, evaluated any previous stimulus. Listen carefully to the first recording. Then listen to the second one and give a score. You can play the audio recordings several times."

The first recording had a standard speech rate - average speed, 7 pauses, and average pause duration. Participants then rated the tempo of the second recording relative to the first. They listened to the standard recording every time before a new one. The experiment was divided into three parts, each presenting various stimuli with different articulation rates, pause numbers and durations. The order of recordings was randomized for each participant.

A preliminary test with seven participants confirmed the clarity of the task, with no reported difficulties. Participants also provided information on their age, gender, native language, and musical training.

3.2 Results of the Experiment with the Female Speaker

The first stage of the experiment focused on the female speaker. It involved 10 participants (6 women and 4 men, aged 19 to 23). Figure 2 illustrates the distribution of responses for each participant. The vertical axis represents participant ratings, and the horizontal axis represents recordings from the fastest (highest speech rate, fewest short pauses) to the slowest. Each group of recordings with the same articulation rate is highlighted in a different color.

The minimum rating was 1, given twice by the same participant, while the maximum rating was 40, given once. Interestingly, recordings with objectively the highest or lowest articulation rates were not always rated as the fastest or slowest ones. It is evident that within groups of equal articulation rates, ratings decreased with increasing pause duration. In some cases recordings with lower articulation rates were rated faster due to the use of pauses. Nonetheless, the graphs show that responses from all participants generally decrease with increasing articulation rate or decreasing pause duration. In the comparison of two identical stimuli (the same recording presented twice), 7 out of 10 participants rated it as 10, while the other three rated it higher, giving 11, 13, and 15.

The recording with the closest average rating to 10, with the least variation (seven ratings of 10, two of 11, and one of 10.5), was the one with a medium speech articulation and 7 pauses, but with shorter pause durations than the standard. This suggests that participants tend to rate speech tempo as faster than it actually is.

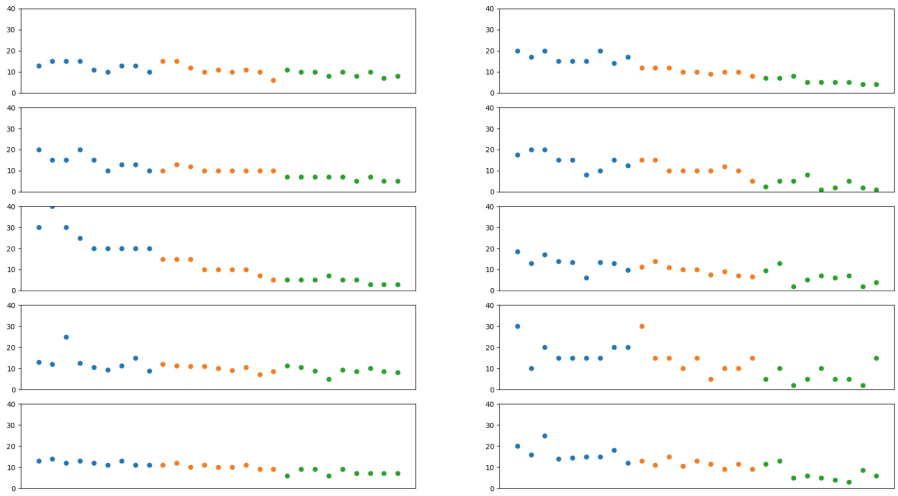


Fig. 2. Distribution of speech rate ratings by each participant for the female speaker. Blue group represents recordings with fast articulation rate, orange - medium, green - slow. Within each group, audio recordings are arranged by increasing total pause duration. (Color figure online)

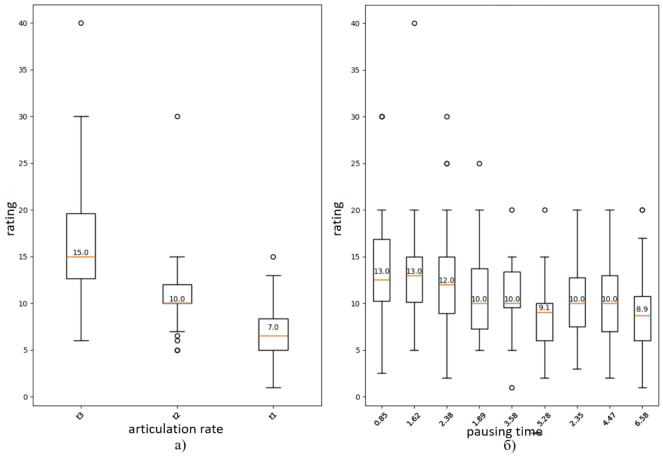


Fig. 3. Ratings grouped by a) articulation rate, b) total pause duration for the female speaker.

Figure 3 presents the results grouped by articulation rate (3a) and total pause duration (number of pauses × duration) (3b). The data indicate that articulation rate had a more significant impact on participant ratings, compared to pause

duration. However, single outliers, such as in the second box in Fig. 3a, show that shorter pause durations could compensate for medium articulation rates, making the overall tempo perceived as faster.

To test the model presented by Lane and Grosjean [12] with the data received, we used the least squares method. The average rating for each recording was calculated, and the results were analyzed using Excel. The R^2 coefficient was 0.73, indicating a moderate fit. Thus, the proposed model may not be entirely suitable for the Russian language.

The model coefficients were recalculated in order to fit the data. Speech rate and total pause duration were normalized per word, resulting in average articulation times from 308 to 369 ms per word, and pause times from 61 to 168 ms. Average ratings were then calculated for three speech rate levels and nine pause levels.

Using Python, the data was plotted on a logarithmic scale, and a least squares approximation line was fitted (Fig. 4). The resulting coefficients were -0.22 for pauses and -5.3 for articulation rate, yielding the model:

$$E' = 0.03A^{-5.3} + 6.02P^{-0.22} - 10 \tag{2}$$

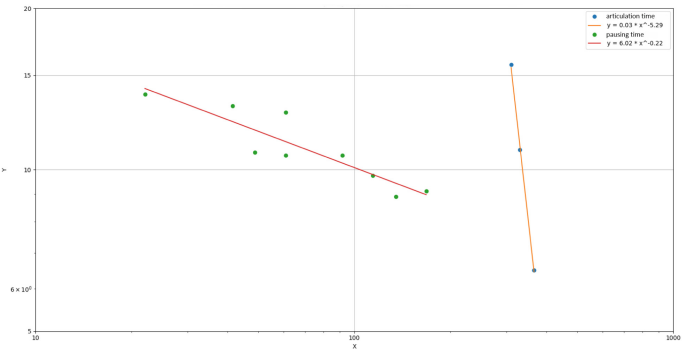


Fig. 4. Speech tempo perception as a function of articulation time (orange line) and pause time (red line) for the female speaker. (Color figure online)

Figure 5 shows the relations between the results predicted by two models and those which were obtained in the experiment. This model had an R^2 of 0.9, a significant improvement over Lane and Grosjean's formula (1) with R^2=0.7. This discrepancy could be due to various factors, including differences in measurement units (words vs. syllables) and the specific characteristics of the Russian language.

The findings suggest that to neutralize articulation rate (i.e., make fast speech appear slow), pause duration must be significantly increased. Conversely, to make speech with many pauses sound fast, a smaller increase in articulation rate is required.

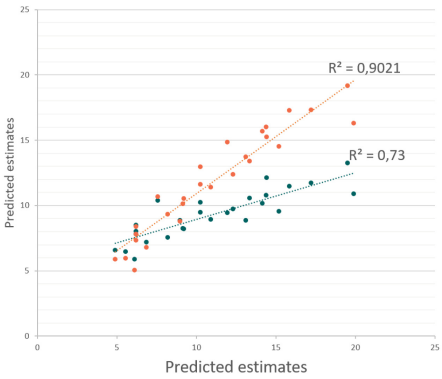


Fig. 5. Relation between rate estimates predicted by a multilinear model and those obtained in the experiment (green corresponds to the initial model, orange - the one that was recalculated) for the female speaker. (Color figure online)

3.3 Results of the Experiment with the Male Speaker

The second stage of the experiment involved the male speaker, with the same ten participants (4 men and 6 women, aged 19 to 23). The distribution of responses appears to be similar to the previous experiment, with ratings ranging from 1 to 40 (Fig. 6). In the task with identical stimuli, 4 out of 10 participants rated it higher than 10: 11, 12, 17, and 20. We observed the similar distribution in the previous experiment but with less variation.

The recording with the closest average rating to 10 and the least variation was the one with a medium articulation rate and stimulus-equivalent pause duration, but with the maximum number of pauses. Interestingly, the maximum rating (40) was given not only to the fastest recording but also to a recording with the average articulation rate and the minimum number of maximum-duration pauses. This suggests that the participant made a judgment within the first few seconds, unaffected by pause duration.

Overall, we did not observe any great differences in the rating distributions between the two speakers. However, participants independently determined their rating boundaries, leading to varied patterns-some with more spread out ratings and others with a straight line pattern (Fig. 2 and 6).

The median rating for medium articulation rate, as shown in Fig. 7a, was 11, with the spread not significantly different from the high rate, unlike for the female voice. This could be partly explained by the smaller difference in articulation rate (7.3% for the male voice vs. 10% for the female voice). The average rating for the slowest articulation rate was 2 points lower. Using this data, Lane and Grosjean's model [12] yielded an R^2 of 0.69, while the model from the female speaker's data yielded R^2 of 0.81, a better result.

The articulation times for the male speaker ranged from 310 to 334 ms per word, and pause times from 15 to 128 ms. Based on average ratings for 3 speech rate levels and 9 pause levels, the data was plotted on a logarithmic scale, and a

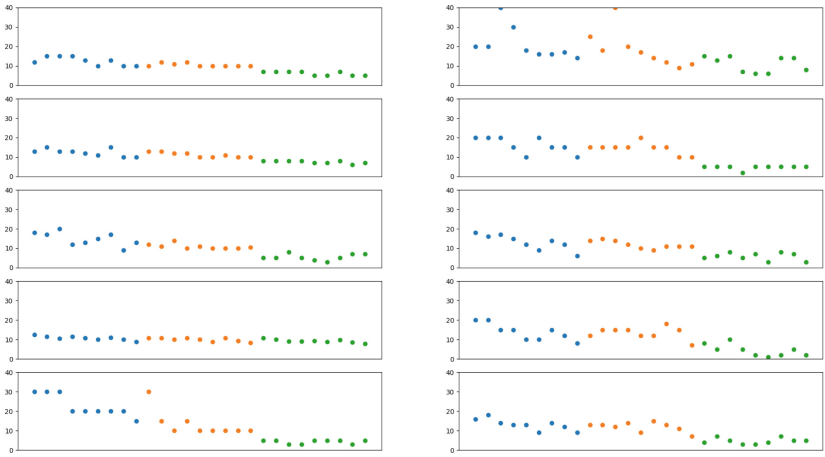


Fig. 6. Distribution of speech rate ratings by each participant for the male speaker. Blue group represents recordings with fast articulation rate, orange - medium, green - slow. Within each group, audio recordings are arranged by increasing total pause duration. (Color figure online)

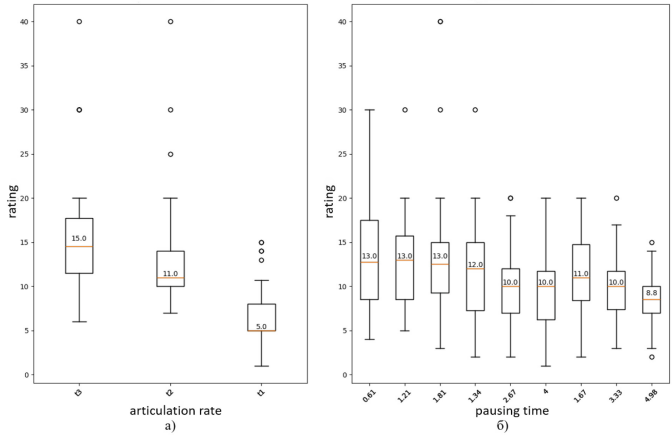


Fig. 7. Ratings grouped by a) articulation rate, b) total pause duration for the male speaker.

least squares approximation line was fitted (Fig. 8), then the model was derived as:

$$E' = 0.02A^{-5.61} + 5.6P^{-0.23} - 10 \tag{3}$$

As shown in Fig. 9, this model explained the data with an R^2 of 0.87. The pause influence coefficient was similar to previous experiments, while the articulation rate exponent was close but differed more from the original model.

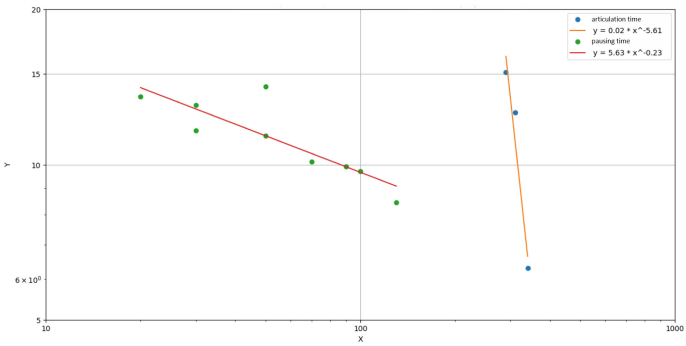


Fig. 8. Speech tempo perception as a function of articulation time (orange line) and pause time (red line) for the male speaker. (Color figure online)

3.4 Comparison of Results

The two experiments showed no significant differences in ratings for male or female voices. Despite the male voice having a higher articulation rate with fewer pauses, the relative scales allowed the comparison. In most cases (18 out of 27 for the female speaker and 17 for the male speaker), participants clearly determined that the recording was faster or equal to the stimulus, or slower or equal. Ambiguous ratings were given to recordings representing extremes (fastest with longest pauses and vice versa). This indicates that for some participants, speech rate was the decisive factor, while for others, it was pauses. We did not find any clear pattern in rating distribution. The obtained models are in many respects similar and equally different from the one proposed by Grosjean and Lane [12]. Combining the results from the first and second experiments allows us to further refine the coefficients, the model takes the form:

$$E' = 0.09A^{-4.2} + 5.9P^{-0.22} - 10 \tag{4}$$

Obviously, extending the range of input data and increasing the number of subjects would further refine the coefficients needed to mathematically describe speech rate perception.

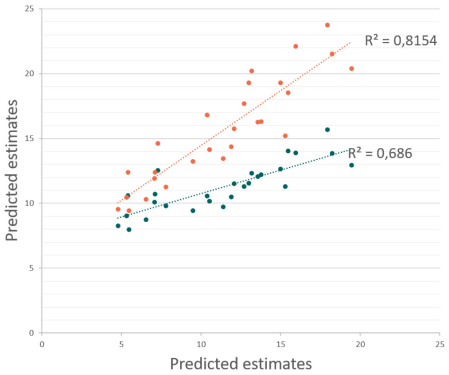


Fig. 9. Relation between rate estimates predicted by a multilinear model and those obtained in the experiment (green corresponds to the initial model, orange - the one that was recalculated) for the male speaker. (Color figure online)

4 Discussion and Conclusion

This study explored the perception of speech tempo, focusing on factors such as speaker and listener gender, articulation rate, and pauses. Our experiments, conducted with recordings of Russian text readings, yielded significant insights into these aspects.

Key findings indicate that male voices are generally perceived as faster, consistent with prior research [4]. This perception disparity did not extend to overall speech tempo, suggesting that pause duration may play a more crucial role. Listener gender showed no significant effect, although men found it more challenging to judge tempos when hesitation and filled pauses were present.

The application of Stevens' magnitude estimation method was effective, with participants consistently overestimating speech tempo. This aligns with the idea that speech rate perception can be described through its components, such as pause duration and articulation rate. The pause coefficient closely matched previous studies by Lane and Grosjean [12], while the articulation rate coefficient did not, possibly due to language system differences or external factors affecting the experimental material.

Despite these variances, coefficients obtained from experiments with two speakers were relatively consistent. The data suggest that while individual factors influence perception, overarching patterns remain identifiable.

Future research should expand the dataset and explore spontaneous speech contexts to further refine the models. These findings enhance our understanding of speech tempo perception and underscore the importance of considering both linguistic and social factors.

References

1. Anderson, N.H.: Algebraic models in perception. Handbook Percept. **2**, 215–298 (1974)
2. Baddeley, A.D.: Time-estimation at reduced body temperature. Am. J. Psychol. **79**, 475–479 (1966)
3. Baron, R.M., Kenny, D.A.: The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. J. Pers. Soc. Psychol. **51**(6), 1173–1182 (1986)
4. Biemans, M.: Gender variation in voice quality. LOT, Utrecht, Netherlands (2000)
5. Bond, R.N., Simpson, S., Feldstein, S.: Relative and absolute judgments of speech rate from masked and content-standard stimuli the influence of vocal frequency and intensity. Hum. Commun. Res. **14**(4), 548–568 (1988)
6. Bosker, H. R.: Our own speech rate influences speech perception. In: Speech Prosody, pp. 227–231 (2016)
7. Chistovich, L. A., Ventsov, A. V., Granstrem, M. P.: Speech physiology. In: Human Speech Perception. Nauka (1976). (In Russian)
8. Eefting, W., Rietveld, A.C.M.: Just noticeable differences of articulation rate at sentence level. Speech Commun. **8**(4), 355–361 (1989)
9. Feldstein, S., Dohm, F.A., Crown, C.L.: Gender as a mediator in the perception of speech rate. Bull. Psychon. Soc. **31**, 521–524 (1993)
10. Grosjean, F.H., Lass, N.J.: Some factors affecting the listener's perception of reading rate in English and French. Lang. Speech **20**(3), 198–208 (1977)
11. Grosjean, F., Lane, H.: Effects of two temporal variables on the listener's perception of reading rate. J. Exp. Psychol. **102**(5), 893 (1974)
12. Grosjean, F., Lane, H.: How the listener integrates the components of speaking rate. J. Exp. Psychol. Hum. Percept. Perform. **2**(4), 538–543 (1976)
13. Kachkovskaia, T., Kocharov, D., Skrelin, P., Volskaya, N.: CoRuSS - a new prosodically annotated corpus of russian spontaneous speech. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), pp. 1949–1954 (2016)
14. Koreman, J.: Perceived speech rate: the effects of articulation rate and speaking style in spontaneous speech. J. Acoust. Soc. Am. **119**(1), 582–596 (2006)
15. Lane, H., Grosjean, F.: Perception of reading rate by speakers and listeners. J. Exp. Psychol. **97**(2), 141–147 (1973)
16. Maslov, S. Yu.: Introduction to Linguistics, 2nd edn. Higher School (1987). (In Russian)
17. McFadden, D.: Sex differences in the auditory system. In: Gonadal Hormones and Sex Differences in Behavior, pp. 261–298. Psychology Press (2014)
18. Miller, J.L., Grosjean, F.: How the components of speaking rate influence perception of phonetic segments. J. Exp. Psychol. Hum. Percept. Perform. **7**(1), 208 (1981)
19. Neff, D.L., Kessler, C.J., Dethlefs, T.M.: Sex differences in simultaneous masking with random-frequency maskers. J. Acoust. Soc. Am. **100**(4), 2547–2550 (1996)
20. Nikandrov, V.V.: Psychophysics and Psychophysical Methods. Rech, SPb (2005). (In Russian)
21. Ornstein, R.E.: On the Experience of Time. Penguin Books, Baltimore (1969)
22. Putman, W.B., Street Jr, R.L.: The conception and perception of noncontent speech performance: implications for speech-accommodation theory (1984)

23. Quené, H.: On the just noticeable difference for tempo in speech. J. Phon. **35**(3), 353–362 (2007)
24. Schwab, S.: Relationship between speech rate perceived and produced by the listener. Phonetica **68**(4), 243–255 (2012)
25. Shiffman, X.R.: Sensation and Perception, 5th edn. Piter, SPb (2003). (In Russian)
26. Stevens, S.S.: The direct estimation of sensory magnitudes: loudness. Am. J. Psychol. **69**(1), 1–25 (1956)
27. Tauroza, S., Allison, D.: Speech rates in British English. Appl. Linguis. **11**, 90–115 (1990)
28. Trouvain, J.: Tempo variation in speech production. In: Implication for Speech Synthesis. Dissertation, Saarbrücken (2003)
29. Vinogradova, Y.S., Prokaeva, V.O., Riekhakainen, E.I.: There are different kinds of pauses: multidimensional classification of pauses for annotating corpora of russian spoken speech. Russ. Speech **6**, 7–23 (2023). (In Russian)
30. Zabrodin, Yu.M., Lebedev, A.N.: Psychophysiology and Psychophysics, p. 288 (1977). (In Russian)

Human and Machine Keyphrase Perception in Russian Text and Speech

Daria Guseva[1(✉)], Olga Mitrofanova[1], and Mikhail Dolgushin[2]

[1] Saint Petersburg State University, Universitetskaya Embankment, 7-9, St. Petersburg 199034, Russia
{daria.guseva,o.mitrofanova}@spbu.ru
[2] St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), 14th Line, 39, St. Petersburg 199178, Russia
dolgushin.m@iias.spb.su

Abstract. The article examines the perception and extraction of keyphrases in both written and spoken text. Experiments were performed on the dataset including transcripts and audio recordings of lectures by Russian-speaking participants of the project "Postnauka". The results show that automated methods for keyphrase extraction have limited accuracy, with statistical algorithms performing the worst and generative AI models, such as ChatGPT, showing a closer resemblance to human perception. Additionally, while there is some overlap between keyphrases extracted from written and oral texts, spoken text presents greater variability. Experiments using synthesized speech indicate that listeners rely heavily on content, rather than acoustic cues, when understanding spoken text. Acoustic analysis reveals that keyphrases are distinguished by longer duration, wider pitch range, and higher energy, aligning with previous findings in other languages.

Keywords: Keyphrase Extraction · Perception · Russian Language · Expert Annotation · Acoustic Analysis

1 Introduction

This research presents a comparative analysis of perceptual experiments designed to study keyphrases in written and spoken texts. Experiments were conducted on a dataset comprising transcripts and audio recordings of lectures delivered by Russian-speaking participants of the project "Postnauka". Based on this material, several hypotheses were tested.

We hypothesized that human perception of oral and written text differs in terms of keyphrase extraction. Additionally, we expected that manually extracted keyphrases would differ from those obtained using various automated methods. Furthermore, we investigated the hypothesis of a prosodic specificity of keyphrases. The article presents the results of our research on keyphrases, examining both their perception and acoustics.

In modern linguistics, communication is viewed as a multimodal multi-level process that uses various modes of expression, with speech communication serving as the

A. Karpov and V. Delić (Eds.): SPECOM 2024, LNAI 15299, pp. 265–280, 2025.
https://doi.org/10.1007/978-3-031-77961-9_20

foundation for verbal interaction between people. The process of transmitting information can be performed through oral and written speech, which are not considered equal. Oral speech is historically primary, meaning it existed before written language. Written language emerged as a way to record and preserve oral language. Oral speech allows for the use of intonation, pausing, and other forms of paralanguage that are absent in written communication.

Perception of the content of written and oral text differs, including at the level of highlighting keyphrases. Manual keyphrase extraction is a labor-intensive task, making automation essential for large datasets.

The increasing volume of multimedia data (videos, music, podcasts, lectures, etc.) consumed by individuals in their daily lives necessitates efficient approaches for indexing and browsing these multimedia documents. As such data may contain audio information describing the main content, automatic keyphrase extraction from speech has become increasingly relevant [1].

Automatic Speech Summarization (SSum) finds wide application in various domains, including extracting vital information from news broadcasts [2], podcasts [3], clinical conversations [4], and business meetings [5]. Notably, keyphrase extraction from oral text often serves as a crucial step within SSum techniques.

Keyphrases in spoken text are characterized by a certain prosodic specificity, which creates the foundations for comparison of the strategies for extracting keyphrases in written and oral text. To the best of our knowledge, the association between keyphrases and prosodic features remains relatively unexplored for the Russian language. A significant amount of research has focused on keyphrase extraction from text, but far fewer studies have been conducted on spoken language data. Even fewer studies utilize information from audio signals, such as prosodic features.

Within this study, both written and oral texts were examined, the keyphrases in which were identified manually and automatically. This approach was chosen to determine algorithms that would better replicate the mechanisms of selecting keyphrases by native speakers.

2 Related Work

Keyphrases are considered in various scientific fields, including computer and cognitive linguistics, psycholinguistics, computer science, information retrieval, philological studies, and others. Keyphrases carry the most important information about the text, presenting it in a compressed format, and contribute to the structuring, classification, summarization, and quick assessment of document content [6, 7]. This suggests that keyphrases directly participate in the process of text perception.

Studies have shown keyphrases are prosodically distinct. Research on spoken Russian has revealed that pauses are less likely to occur before keyphrases compared to other words but more frequent after keyphrases [8]. Keyphrases can also be characterized by a longer duration, a wider pitch range and higher energy. However, this research is limited to English and Mandarin Chinese, further investigation is required for other languages [1].

It is also worth mentioning the concept of a prosodically highlighted word. This refers to words receiving accentual emphasis, signifying their perceptual prominence

within a phrase through prosodic means [9]. Keyphrases may, but do not necessarily, be prosodically highlighted [10].

As previously stated, the task of keyphrase extraction from speech often arises as a component of Speech Summarization (SSum), where the goal is to generate a concise summary of spoken content, either in the form of speech or text, while preserving core meaning and avoiding loss of crucial information [11].

SSum models can be categorized into transcript-based and acoustic-based. Transcript-based models utilize Automatic Speech Recognition (ASR) to transcribe the audio, then apply text summarization techniques [12, 13]. Consequently, keyphrases are extracted from the recognized text data, and their identification employs the same methods applicable to extracting keyphrases from written text.

While transcript-based methods benefit from advancements in text summarization, their accuracy depends on the availability and quality of ASR, which can be unreliable in noisy environments or with low-resource languages. Such models also suffer from the absence of acoustic information conveying the speaker's emotions.

Approaches to acoustic-based speech summarization are less prevalent and remain under-explored in the scientific literature. One of the recent proposed approaches is E2E (end-to-end) Sum which generates abstractive summaries directly from speech, bypassing ASR errors and utilizing acoustic information. Despite promising results, such models require substantial training data and tend to produce unnatural sentences in data-scarce scenarios [12].

Current three-stage method, ESSumm (extractive speech-to-speech summarization), avoids the need for ASR entirely [11]. It utilizes a pre-trained Wav2Vec2.0 model to extract audio features, employs k-means clustering for phoneme probability representation, and leverages Latent Semantic Analysis to assess segment importance. This approach enables real-time summarization of spontaneous conversations and potential output generation in audio format.

Therefore, the task of automatic keyphrase extraction from speech is primarily viewed as a component within speech summarization and, depending on the summarization model, can be implemented either through keyphrase extraction methods from written text or by leveraging acoustic signal processing.

3 Experimental Dataset

The experimental dataset developed for our study includes audio recordings of speech by Russian-speaking lecturers from the project "Postnauka"[1]. Each audio recording presented within the project is accompanied by a textual transcript. For the case discussed in this paper, we selected two audio recordings of male[2] and female[3] speakers. Narrators are professional lecturers who work with audiences of different ages. The topics of their short lectures deal with general linguistics. Accordingly, in the first lecture, the existence of grammatical genes is discussed, along with the concept of universal grammar, the

[1] https://postnauka.ru.

[2] https://postnauka.org/video/57524.

[3] https://postnauka.org/video/61500.

principle of finding structure in chaos, and the acquisition of new words by children. The second lecture is dedicated to lexical collocations, in particular, to bag of words models and logarithmic transformations in linguistics.

The choice of the given dataset is due to the fact that lectures belong to the popular science style and are intended for a wide audience of non-specialists, which allows for involvement of informants in keyphrase annotation irrespective of their professional specialization.

Readability of the texts was evaluated using several metrics adapted for the Russian language[4]. Both texts have approximately the same level of complexity, comparable to that of a high school student, college student, or university graduate.

Audio recordings synthesized from the transcripts of the lectures were also used as material. The Free Text to Speech tool[5], which is freely available and utilizes the Microsoft AI Speech Library, was employed to generate the synthesized samples. The choice of this tool was based on its accessibility. The service requires no prior registration and allows for an unlimited number of speech synthesis from text, as well as enables the downloading of the resulting audio recording in MP3 format. Two neutral voices (male and female) for Russian language synthesis are available, as well as the option to customize additional parameters, which were not utilized in this study.

The first text, corresponding to the audio recording of the female speaker, consists of 8 paragraphs (1.191 graphic word). The second text, from the male speaker, has 10 paragraphs (1.436 graphic word). The audio version of the first text lasts 12 min, while the second text is 13 min in duration. Each audio recording is accompanied by manual annotations for phrases and words. The first synthesized audio recording has a duration of 10 min. The second audio recording lasts 12 min.

4 Experimental Procedure

During the study, keyphrases were extracted automatically and manually by the participants of the perceptual experiments based on written and oral texts (either natural or synthesized). Table 1 shows parameters of the experiments.

Additionally, acoustic characteristics of the keyphrases were calculated. For this analysis, the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) was employed [14]. For further details on this feature set, refer to Sect. 5.4.

In this study, the following algorithms were used for automatic extraction of keyphrases in written text: (1) statistical (Chi-square, Log-Likelihood, PMI-test, T-test), (2) hybrid linguostatistical (RAKE, RuTermExtract, SpaCy), (3) machine learning-based methods (KeyBERT). The list of methods is not limited to those mentioned above. In this study, algorithmic implementations of statistical methods from the NLTK library were used.

Before the procedure of keyphrase extraction the text was preprocessed to represent it in a format suitable for subsequent recognition: tokenization (identifying word forms in the text) and removal of stop words. It is worth noting that stop words are discourse

[4] https://github.com/SergeyShk/ruTS.
[5] https://www.text-to-speech.online.

Table 1. Experimental parameters.

Parameter	Automatic keyphrase extraction	Expert annotation
Functional style of the text	Popular science style	Popular science style
Functional register of the text	Written	Written and oral
Length of keyphrases	Limitations depend on the method	Unigram, bigram, or trigram
Size of the list of keyphrases	1…10	1…10
Keyphrase ranking method	In descending order of importance	In descending order of importance

markers that depend on specific texts and tasks. An attempt was also made to lemmatize the texts, but lemmatization worsened the results achieved for our dataset as it restricted the output.

Keyphrases were also extracted using ChatGPT, a neural language model based on artificial intelligence developed by OpenAI. These results were considered as a relative benchmark in subsequent stages when comparing the results obtained automatically and through perceptual experiments.

The top 10 results from the keyphrase extraction methods were analyzed.

Manual annotation was obtained as a result of a series of perceptual experiments involving Russian-speaking participants (readers and listeners). The oral texts were presented as both natural and synthesized speech samples. The Google Forms platform was used to create and conduct the research.

Participants in the experiment, based on written text material, were required to identify 10 keyphrases from the text after reading, which convey the main content of the document, and rank them from most important to less important. In the experiment using oral text material, listeners performed a similar task after listening to an audio recording. The surveys were designed in such a way that participants had the opportunity to read the text or listen to audio recordings several times.

Table 2 provides information about the participants.

It is known that the number of keyphrases in a set can vary within a fairly wide range, but for optimal processing of oral text, a small set of keyphrases is preferred [6, 9]. Therefore, the allocation of exactly 10 keyphrases allows for some flexibility in terms of relevance.

The instructions specified a restriction on the type of phrases that could be chosen as keyphrases – participants could select unigrams, bigrams, or trigrams (keyphrases consisting of one, two, or three words, respectively).

The results of keyphrase extraction were also compared to annotations for prosodically highlighted words, obtained from 5 expert phoneticians. Keyphrase sets and prosodically highlighted word sets were generated independently for the same texts. The experts were required to listen to the audio recordings and mark the prosodically highlighted words in the text while listening.

Table 2. Information about the participants.

Material on which the experiment is based	Number of the participants	Average age of the participants
First and second written texts	69 (49 female, 20 male)	25
First natural oral text	50 (32 female, 18 male)	21
Second natural oral text	52 (30 female, 22 male)	23
First synthesized text	50 (33 female, 17 male)	27
Second synthesized text	48 (26 female, 220 male)	29

5 Results and Discussion

5.1 Keyphrase Extraction Results Using Automatic Methods on Written Text and Manually from Written and Natural Oral Text

When applying automated methods, statistical methods were clustered based on the proximity of their keyphrase extraction results. Among the methods discussed, Chi-square - PMI-test and Log-Likelihood - T-test pairs exhibited close groupings. Similar keyphrase groups ultimately yielded pairs of statistical methods, as well as hybrid linguostatistical methods RuTermExtract and SpaCy. A similar pattern was observed when working with both texts.

The keyphrase sets extracted using various automatic methods demonstrated limited overlap. To evaluate their alignment with actual human perception, experiments were conducted with participants.

We compared the results of keyphrase extraction using automatic methods in written text with those manually extracted by participants in experiments using both written and natural oral texts. The complete list of manually extracted keyphrases was reviewed for spelling errors and typos.

Table 3 summarizes the overlap between keyphrases identified by participants and those extracted by various automatic algorithms for different text types (written and oral). It shows (1) the number of total responses, (2) the number of unique keyphrases identified in each text type, (3) the top 10 most frequently identified keyphrases in each text, (4) the percentage of responses accounted for by those top 10 keyphrases, (5) which algorithms successfully identified some of the same keyphrases as participants.

It is also noteworthy that while the keyphrase sets identified by participants for the same text presented in written and spoken forms exhibited a degree of overlap, they demonstrated different relative frequency patterns.

Table 4 compares keyphrase extraction in written and oral texts. It shows that oral texts generally resulted in lower relative frequencies of keyphrase extraction and fewer keyphrases achieving a high frequency. The relative frequency of keyphrase extraction was lower for the oral texts. For example, the keyphrase “универсальная грамматика” (“universal grammar”), which ranked second in both sets (text 1), had a relative frequency of 0.080 for the written text and 0.069 for the oral text.

Table 3. Manual keyphrase extraction overlap with automatic methods.

Experiment	Total responses	Unique keyphrases	Top 10 keyphrases	% the top 10 of total responses	Overlapping automatic methods (number of matches)
Written text 1	660	137	341	51.67%	Chi-square, PMI-test (1); Rake, KeyBERT, SpaCy (2); RuTermExtract (4); ChatGPT (6)
Written text 2	660	157	307	46.52%	Log-likelihood, T-test (1); KeyBERT, SpaCy (2); RuTermExtract; ChatGPT (4)
Oral text 1	490	166	198	40.41%	Chi-square, PMI-test, Rake (1); KeyBERT, SpaCy (2); RuTermExtract (5); ChatGPT (6)
Oral text 2	490	189	146	29.8%	Log-likelihood, T-test, Rake, KeyBERT (1); RuTermExtract (2); ChatGPT (3)

Table 4. Comparison of keyphrase extraction results in written and oral texts.

Text	Number of overlapping keyphrases	Relative frequency of the keyphrases	Keyphrases with frequency > 0.05
Text 1 (written & oral)	7	Oral $<$ written	4 (written) $=$ 4 (oral)
Text 2 (written & oral)	5	Oral $<$ written	4 (written) $>$ 2 (oral)

Based on the results obtained for the second text (greater variability in keyphrase sets; lower relative frequency values for keyphrases compared to the first text), it can be inferred that it presented a greater challenge for participants' comprehension, despite both texts having the same readability level.

When comparing the keyphrase sets generated by various automatic methods and the results of perceptual experiments, statistical algorithms exhibited the least satisfactory

results, with matches to participant responses only observed in isolated cases. This is attributed to the inherent limitations in the applicability of statistical methods. ChatGPT demonstrated the highest number of matches with the responses provided by the participants.

While keyphrase sets extracted from identical written and oral texts exhibit some overlap, they display significant variation in the relative frequencies of keyphrases. This suggests that oral texts tend to produce more diverse keyphrase sets compared to written texts.

5.2 Keyphrase Extraction from Synthesized Texts

The sets of the 10 most frequent keyphrases extracted from both oral texts – natural and synthesized – were compared. The results of the comparison for the first text are presented in Table 5, with matching keyphrases highlighted in bold italics.

Interestingly, listeners of the synthesized text demonstrated greater consistency in choosing keyphrases: the relative frequency for matching keyphrases is higher or approximately the same in the results for the synthesized text as can be clearly seen in Table 5. As with the first text, the relative frequency for keyphrases extracted from the second synthesized text is higher than for analogous keyphrases extracted from the natural oral text. These coincidences suggest that audiences relied on the content level of the texts when selecting keyphrases, while the acoustic component did not exert a significant influence.

Table 5. Comparison of Keyphrase Extraction Results from the First Oral Text – Natural and Synthesized.

Natural text		Synthesized text	
Keyphrase	Relative frequency	Keyphrase	Relative frequency
Грамматические гены (grammatical genes)	0.076	*Грамматические гены (grammatical genes)*	0.086
Универсальная грамматика (universal grammar)	0.069	*Универсальная грамматика (universal grammar)*	0.078
Когнитивный взрыв (cognitive explosion)	0.055	*Когнитивный взрыв (cognitive explosion)*	0.076
Грамматический взрыв (grammatical explosion)	0.053	*Грамматический взрыв (grammatical explosion)*	0.054
Структура (structure)	0.029	Мнемотехника (Mnemonics)	0.048
Структура в хаосе (structure in chaos)	0.029	*Структура (structure)*	0.028
Ищи структуру в хаосе (look for structure in chaos)	0.027	*Структура в хаосе (structure in chaos)*	0.028

(*continued*)

Table 5. (*continued*)

Natural text		Synthesized text	
Keyphrase	Relative frequency	Keyphrase	Relative frequency
Эксперимент (experiment)	0.027	*Эксперимент (experiment)*	0.026
Ноам Хомский (Noam Chomsky)	0.020	Языковые высказывания (language statements)	0.026
Язык (language)	0.020	*Ищи структуру в хаосе (look for structure in chaos)*	0.024

Table 6 compares the results of keyphrase extraction from both spoken texts, synthesized and natural.

Table 6. Comparison of keyphrase extraction results in natural and synthesized texts.

Text	Number of unique keyphrases	Top 10 keyphrases (number of responses)	Percentage of total responses	Overlap with natural text
Synthesized text 1	163	237	47.4%	8 out of 10 keyphrases in the top 10 set
Synthesized text 2	106	241	50.2%	7 out of 10 keyphrases in the top 10 set

For natural oral texts, the ratio of unique keyphrases to the total number of responses was 40.41% for the first text and 29.8% for the second as shown in Table 3. Similar metrics for synthesized texts were higher −47.4% and 50.2%.

Previously, analyzing the second natural text, we hypothesized that despite its readability level comparable to the first text, it might have been more challenging to perceive, as less consistency was observed among listeners when extracting keyphrases. However, since the ratio of unique keyphrases to the total number of keyphrases is 1.5 times higher for the same synthesized text, it is likely that the perception of the synthesized text enhanced the consistency among listeners. Some increase in consistency was also observed for the first text. It can be assumed that listeners were either more attentive when analyzing the content of the synthesized speech or were not influenced by certain characteristics that affected perception in the case of the natural text. However, further analysis on a larger corpus is required to refine this hypothesis.

5.3 Keyphrase Extraction Results and Prosodically Highlighted Words

To test the hypothesis that keyphrases possess a certain prosodic specificity, the results of keyphrase extraction identified by participants from the spoken texts (natural speech samples) were compared with annotations for prosodically highlighted words.

Comparing presents a significant challenge. One issue arises from the fact that certain words and phrases are repeated multiple times within the text. While annotations for prosodically highlighted words allow for precise identification of which repeated word the expert intends, keyphrase extraction does not involve isolating specific segments. Additionally, keyphrases can consist of multiple words, while prosodic highlighting typically applies to individual words, even if they are adjacent within the text structure. These characteristics hinder the direct comparison of such datasets. Due to the above-mentioned reasons, we do not present the results of analyzing the inclusion of keyphrases within the sets of prosodically highlighted words, but instead analyze the inclusion of prosodically highlighted words within the sets of keyphrases.

The auditory analysis conducted by expert phoneticians resulted in a complete set of 350 prosodically highlighted words for the first text and 263 elements for the second text. Since over half of the words in these sets (61% for the first and 52% for the second) were highlighted by only one expert, the decision was made to consider only those words selected by at least two experts.

Table 7 compares prosodically highlighted words with manually extracted keyphrases in two oral texts.

Table 7. Keyphrase Extraction Results & Prosodically Highlighted Words.

Text	Prosodically highlighted word selected by at least 2 experts	Number of extracted keyphrases	Overlap (prosodically highlighting & keyphrases)	Concordance coefficient
Text 1	103	166	27.18%	0.00113
Text 2	169	189	22.94%	0.00187

While there are some overlapping words, less than 30% of the prosodically highlighted words coincided with keyphrases, meaning that a significant portion of the analyzed set of prosodically highlighted words was not included in the set of keyphrases.

This suggests that prosodic highlighting alone is not a strong indicator of keyphrase status, and that other factors likely contribute to human perception of keyphrases. The table also highlights the low agreement between experts regarding prosodic highlighting, as indicated by the concordance coefficient.

5.4 Acoustic Analysis of the Keyphrases

This study investigated the acoustic correlates of keyphrase extraction by human listeners. An acoustic analysis was conducted on words, phrases up to three words, and

sentences in natural oral texts. The aim was to investigate the presence of acoustic correlates associated with the extraction of these units as keyphrases by human listeners.

Acoustic features were extracted using the OpenSMILE library[6] and the eGeMAPS v0.2 feature set. This set was chosen for its efficiency, offering a minimal number of significant features, facilitating rapid feature extraction and simplifying interpretation. The eGeMAPS set encompasses a total of 88 parameters. The parameters for this feature set were selected based on factors such as: (1) their ability to reflect physiological changes in voice production associated with the transmission of emotional states, (2) their proven value in previous research, (3) the availability of automatic extraction methods, and (4) their theoretical significance.

While more comprehensive feature sets, such as ComParE [15], could potentially provide a more complete understanding of the speech processes, eGeMAPS v0.2 was deemed sufficient for this study's objectives.

Analysis revealed weak positive or negative Pearson correlations between certain acoustic features and keyphrase extraction, as presented in Table 5.

Notably, accounting for the position of the keyphrase within a sentence and the duration of pauses before and after keyphrases and other words did not reveal any significant correlation with their extraction by listeners. This is despite the correlation between keyphrases and pauses after them noted in [8]. Further investigation is warranted to understand the discrepancies between these findings.

Table 8 reveals a slight but consistent correlation with spectral and frequency characteristics, suggesting that keyphrases in Russian may be characterized by wider pitch

Table 8. Acoustic analysis of the keyphrases.

Acoustic feature	Meaning of the acoustic feature	Pearson correlation		
		The first text	The second text	Both texts
alphaRatioUV_sma3nz_amean	The arithmetic mean of ratio of the summed energy from 50–1000 Hz and 1–5 kHz	−0.1627	−0.1148	−0.1347
F0semitoneFrom27.5Hz_sma3nz_pctlrange0–2	The logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz (semitone 0). Percentile range of 0 to 20-th percentiles	0.2373	0.1783	0.1847

(*continued*)

[6] https://audeering.github.io/opensmile-python/.

Table 8. (*continued*)

Acoustic feature	Meaning of the acoustic feature	Pearson correlation		
		The first text	The second text	Both texts
F0semitoneFrom27.5Hz_sma3nz_stddevNorm	The logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz (semitone 0) (normalized standard deviation)	0.2563	0.1685	0.1995
F1frequency_sma3nz_stddevNorm	The center frequency of the first formant (coefficient of variation)	0.1307	0.1596	0.1417
F2bandwidth_sma3nz_stddevNorm	The bandwidth of the second formant (coefficient of variation)	0.1359	0.1415	0,1194
F2frequency_sma3nz_stddevNorm	The center frequency of the second formant (coefficient of variation)	0.1151	0.1487	0.1178
F3bandwidth_sma3nz_stddevNorm	The bandwidth of the third formant (coefficient of variation)	0.1413	0.1428	0.1087
F3frequency_sma3nz_stddevNorm	The center frequency of the third formant (coefficient of variation)	0.1509	0.1074	0.1090
hammarbergIndexUV_sma3nz_amean	The ratio of the strongest energy peak in the 0–2 kHz region to the strongest peak in the 2–5 kHz region	0.1685	0.1189	0.1392

(*continued*)

Table 8. (*continued*)

Acoustic feature	Meaning of the acoustic feature	Pearson correlation		
		The first text	The second text	Both texts
jitterLocal_sma3nz_stddevNorm	The deviations in individual consecutive F0 period lengths (coefficient of variation)	0.1785	0.1507	0.1536
loudness_sma3_stddevNorm	The estimate of perceived signal intensity from an auditory spectrum (coefficient of variation)	0.1720	0.1766	0.1756
MeanVoicedSegmentLengthSec	The mean length of continuously voiced regions ($F0 > 0$)	0.1660	0.1080	0.1273
shimmerLocaldB_sma3nz_stddevNorm	The difference of the peak amplitudes of consecutive F0 periods (coefficient of variation)	0.1653	0.1450	0.1466
slopeUV500-1500_sma3nz_amean	The linear regression slope of the logarithmic power spectrum within the two bands (500 and 1500)	−0.1186	−0.1693	−0.1395
spectralFluxV_sma3nz_stddevNorm	The difference of the spectra of two consecutive frames (coefficient of variation)	0.1844	0.2267	0.1921

(*continued*)

Table 8. (*continued*)

Acoustic feature	Meaning of the acoustic feature	Pearson correlation		
		The first text	The second text	Both texts
StddevVoicedSegmentLengthSec	The standard deviation of continuously voiced regions (F0 > 0)	0.17757	0.165537	0.166571

range and higher energy, similar to findings observed in previous research on English and Mandarin [1].

6 Conclusion and Future Work

This study investigated how the perception of written and oral text content differs at the level of keyphrase extraction. A series of experiments were conducted using materials from popular science lectures presented in both written and spoken formats, with the latter employing audio recordings of natural and synthesized speech.

Keyphrases were extracted from written texts using various automated algorithms. To obtain manual annotation of keyphrases, perceptual experiments were organized. In total, 269 auditors participated in the surveys. For spoken texts, annotation of prosodically highlighted words was also obtained, with 5 expert phoneticians involved in the task.

The study yielded several findings.

1. Sets of keyphrases obtained using different automated methods demonstrate little overlap with each other. Statistical algorithms show the lowest effectiveness in keyphrase extraction that coincide with those identified by auditors. The generative AI model ChatGPT shows results closest to actual human perception.
2. Sets of keyphrases extracted from the same written and oral texts partially overlap, but demonstrate differences in relative frequency of keyphrases, indicating greater variability in sets extracted from oral texts.
3. Experiments using neutral synthesized texts indicate that listeners primarily rely on the semantic content of the text when comprehending spoken language, rather than solely on acoustic features. This suggests that automatic identification of keyphrases within the speech stream without relying on text may be a challenging task. Further research with a larger dataset is necessary to confirm this finding.
4. Analysis of data obtained from the perceptive experiment on keyphrase extraction and audio analysis of the original lectures have confirmed that keyphrases might be characterized by longer duration, wider pitch range and higher energy as shown earlier in similar studies on Mandarin Chinese and English.

The results obtained in this study can serve as a foundation for further development of automatic summarization algorithms for spoken texts. Existing methods for summarizing spoken texts primarily focus on keyphrase extraction from recognized textual data. In

this regard, summarization algorithms capable of keyphrase extraction directly from the audio signal are particularly interesting. Developing such models requires considering both the characteristics of human perception of oral texts and the acoustic characteristics of the speech signal.

Acknowledgment. The part of the research related to the assessment of algorithms for automated keyphrase extraction was carried out with support of St. Petersburg State University under the project 124032900006–1.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Chen, Y.-N., Huang, Y., Lee, H.-Y., Lee, L.-S.: Unsupervised two-stage keyword extraction from spoken documents by topic coherence and support vector machine. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5041–5044 (2012)
2. Maskey, S., Hirschberg, J.: Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization. In: Proceedings of the Interspeech 2005, pp. 621–624 (2005). https://doi.org/10.21437/Interspeech.2005-66
3. Vartakavi, A., Garg, A., Rafii, Z.: Audio summarization for podcasts. In: Proceedings of the 2021 29th European Signal Processing Conference (EUSIPCO), Dublin, Ireland, 2021, pp. 431–435 (2021). https://doi.org/10.23919/EUSIPCO54536.2021.9615948
4. Su, J., Zhang, L., Hassanzadeh, H.R., Schaaf, T.: Extract and abstract with BART for clinical notes from doctor-patient conversations. In: Proceedings of the Interspeech 2022, pp. 2488–2492 (2020). https://doi.org/10.21437/Interspeech.2022-10935
5. Riedhammer, K., Favre, B., Hakkani-Tur, D.: Long story short–global unsupervised models for keyphrase based meeting summarization. Speech Commun. **52**(10), 801–815 (2010)
6. Petrova, T.E.: Contextual predictability of text keywords. In: PSU, Perm, pp. 73–77 (2006) (in Russian)
7. Yagunova, E.V.: Variability of strategies for perceiving oral text (experimental study on the material of Russian texts of different functional styles). In: Perm State University, Perm (2008) (in Russian)
8. Yagunova, E.V.: The role of keywords in the perception of sounding and written text (based on the material of the Russian language). In: Proceedings of the International conference on March 14–16, 2002, A person writing and reading: problems and observations, pp. 197–204. Publishing House of St. Petersburg State University, St. Petersburg (2004) (in Russian)
9. Nikolayeva, T.M.: The Semantics of Accentuation, p. 104. Nauka, Moscow (1982) (in Russian)
10. Svetozarova, N.D., Shtern A.S.: Keyphrases and phonetically highlighted words of the text. In: Experimental Phonetics, pp. 157–170. Nauka, Moscow (1989) (in Russian)
11. Wang, J.: ESSumm: extractive speech summarization from untranscribed meeting. In: Proceedings of the Interspeech 2022, pp. 3243–3247 (2022). https://doi.org/10.21437/Interspeech.2022-945

12. Matsuura, K., et al.: Transfer learning from pre-trained language models improves end-to-end speech summarization. In: Proceedings of the Interspeech 2023, pp. 2943–2947 (2023). https://doi.org/10.21437/Interspeech.2023-1307
13. Kotey, S., Dahyot, R., Harte, N.: Query based acoustic summarization for podcasts. In: Proceedings of the Interspeech 2023, pp. 1483–1487 (2023). https://doi.org/10.21437/Interspeech.2023-864
14. Eyben, F., et al.: The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Trans. Affect. Comput. **7**(2), 190–202 (2016). https://doi.org/10.1109/TAFFC.2015.2457417
15. Schuller, B., et al.: The INTERSPEECH 2015 computational paralinguistics challenge: nativeness, Parkinson's & eating condition. In: Proceedings of the Interspeech 2015, pp. 478–482 (2015). https://doi.org/10.21437/Interspeech.2015-179

Assessment of Children's Ability to Manifest Emotions in Facial Expressions, Voice and Speech by Humans, Automatic, and on a Likert Scale

Elena Lyakso[1](✉), Olga Frolova[1], Anton Matveev[1], Aleksandr Nikolaev[1], and Ruban Nersisson[2]

[1] The Child Speech Research Group, St. Petersburg State University, St. Petersburg, Russia
lyakso@gmail.com

[2] School of Electrical Engineering, Vellore Institute of Technology, Vellore, India

Abstract. The goal of the study was to assess children's ability to manifest emotions in facial expressions and speech by humans, automatic and using Likert scale scores. To achieve this goal, two studies were conducted. The first study performed a perceptual and automatic analysis of the emotions "joy - neutral - sadness - anger" in typically developing (TD) children; the second - compared emotion recognition in four groups of children - TD, autism spectrum disorders (ASD), intellectual disabilities (ID) and Down syndrome (DS) by expert and automatic, and analyzed Likert scale scores for completing test tasks. The participants of the study were 110 children aged 5 - 16 years, 18 adults. The original dataset containing video and audio fragments of children's emotional states was used. Experts recognize the emotions of children from all groups by video and speech more accurately than automatic classifications, with higher UAR values for TD children by audio and video in a perceptual experiment and by audio in the automatic classification of emotions. Differences in the classification accuracy of emotions in children with ASD, ID, and DS were identified. Sadness and anger states are automatically classified poorly by audio and video in children with ASD, ID, and DS. The novelty of the results lays in the obtaining normative data on the recognition of emotions in TD children and in comparative data on groups of TD children, children with ASD, ID, DS.

Keywords: Emotional State · Perceptual and Automatic Recognition · Children · Video · Audio Modalities · Likert Scale Scores

1 Introduction

In recent decades, the use of artificial intelligence and neurotechnologies in medicine has been an actively developing area. A special place in medical practice is occupied by the development of technologies for early diagnosis, rehabilitation, training, and the creation of assistive and alternative communication systems using artificial intelligence. Conversational agents, virtual assistants, and educational robots are developed for adults

A. Karpov and V. Delić (Eds.): SPECOM 2024, LNAI 15299, pp. 281–294, 2025.
https://doi.org/10.1007/978-3-031-77961-9_21

and children with Down syndrome (DS) [1, 2], for children with autism spectrum disorders (ASD) [3, 4] and intellectual disabilities (ID) [5]. Assessing the development of the emotional sphere of children, identifying specific features of the manifestation, regulation, and recognition of emotions is one of the important diagnostic components.

The development of the emotional sphere of typically development (TD) children is studied on the material of different cultures [6–8]; the specifics of the formation of different levels of the emotional development in children, depending on age, was described [9]. Educational programs, books, movies, cartoons, and gadgets are aimed at the age of children. For children with atypical development or developmental disorders, the situation is more complicated. Each disease has its own development profile, including the emotional sphere of the child. In the case of developmental disorders and/or atypical development of the child, there may be a discrepancy between the internal state and the external manifestation of emotions [10], a discrepancy in the manifestation of emotions in facial expressions and voice [11].

Individuals with ASD tend to experience difficulties with emotion regulation [12, 13]. It was noted, that there are significant and broad-ranging deficits in emotion processing in ASD, present across a range of stimulus domains and in the auditory and visual modalities [14]. ASD is characterized by impairments in the recognition of emotional and non-emotional facial features, as well as in the recognition of emotions in faces and other modalities [15]. The emotions of children with ID have not been studied sufficiently [16]. The focus is on the needs and emotions of children with severe disabilities [17]. Many works are devoted to studying the recognition of emotions by adults and children with Down syndrome (DS) through facial expression [18], by facial expression and vocalizations [19]. Studies on the manifestation of emotional states by children with DS are few [20, 21], which may be due to traditional views that perceive people with DS as friendly, sociable, charming individuals with a positive mood [22]. The studies for children's emotion classification by voice and facial expressions using the original dataset obtained when performing standardized test tasks by children with developmental disorders are absent.

The methodological approach Child Emotional Development Method (CEDM) developed for assessing the emotional sphere of children allowed us to obtain data on the reflection of emotions in the facial expression, voice and speech of TD children, children with ASD, ID, and DS, to create a dataset and obtain scores for the test tasks completed by children.

The purpose of the study was to assess children's ability to manifest emotions in facial expressions and speech by humans, automatic, and using Likert scale scores.

2 Methods

2.1 Experimental Design

To achieve this goal, two studies were conducted. In the first study, a perceptual and automatic analysis of the emotions "joy - neutral - sadness - anger" of TD children was carried out according to the approaches we used assessing emotions by audio and video in children with DS [23], ASD, and ID [24]. The second study was made to compare the recognition of emotions in children of four groups - TD, ASD, ID, and DS - by experts and

automatically. Testing of children was carried out according to a standardized protocol of the CEDM [25] and to assess the recognition of children's emotions by facial expressions and speech, recordings of identical test tasks performed by all children were used. The implementation of test tasks was assessed on a Likert scale [26]. Therefore, along with the assessment of children's ability to manifest emotions in facial expressions, voice and speech by humans and automatically, Likert scale points were used (Fig. 1).

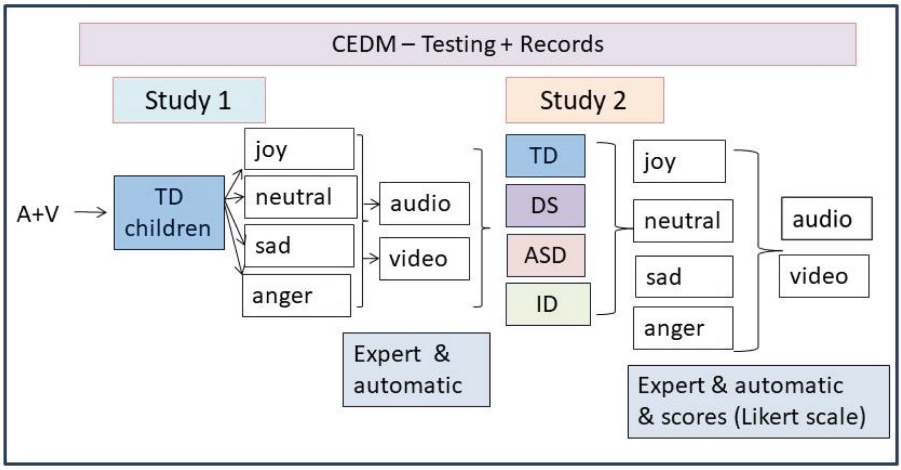


Fig. 1. The design of the experiment.

2.2 Participants of the Study

The participants of the study were 110 children aged 5 - 16 years: 25 children with ASD, 25 children with ID, 35 children with DS, and 35 TD children. The choice of children was carried out in accordance with the selection criteria for testing by CEDM [25]. The number of children in age groups (5–7 years, 8–11 years, 12–14 years, 15–16 years) was the same for each group of participants. The CARS scale, which indicates the severity of autistic disorders, was completed by the parents of children with ASD and ID (autistic disorders may accompany intellectual disabilities as concomitant symptoms); for ASD children scores – 30 - 43 points, mild to moderate severity of autistic disorders, for children with ID - mild to moderate impairment. 14 adults – experts (age 36.7 $\pm$ 13 y; 7 - male, 7 - female) participated in the study, 4 other experts analyzed dataset.

2.3 Data Collection

Video and audio recordings of emotional expression were made in the laboratory condition for TD children, in the Medical Center (St. Petersburg, Russia) (for ASD & ID children), and in the child center of the public organization "Down Center" (St. Petersburg) (for children with DS) in a testing of the Child Emotional Development Method

(CEDM) [25]. The duration of testing was from 1 to 2 h, which was determined by the psychoneurological state of the children and the characteristics of their behavior. The testing scheme was constant [25] regardless of the group of children. Children with ASD, ID, DS were tested in the presence of their parents.

The recording was carried out in a room measuring 18 m^2 - 25 m^2, without a special noise-absorbing wall covering, but with soft puzzle mats on the floor.

For video recording of facial expression of children, a SONY HDR-CX560 video camera (maximum resolution 1920 × 1080 at 50 frames per second) was used, which was located at a distance of 1 m from the child's face. To record children's speech, Marantz PMD660 tape recorder with a SENNHEIZER e835S external microphone was used. The microphone was set at a distance of 30–50 cm from the child's face. Audio files were saved in.wav format, 48000 Hz, 16 bits.

The parents of the children participating in the study signed an informed consent approved by the Ethics Committee of St. Petersburg State University.

2.4 Dataset

The original dataset containing video and audio fragments of children's emotional states was used. From the records of testing children according to the CEDM, two specialists selected the video fragments during which the child demonstrates facial expressions corresponding to one of the four emotional states (joy, neutral, sadness, anger). Only those video clips were selected when the child's face and entire head were completely in the frame and were not covered by hands or toys. All video fragments were annotated by two experts in the state of "joy- neutral (calm) - sadness - anger". We selected only video clips for which the agreement between experts was 1.0 [27]. The audio fragments contained children's speech (the segments with the speech of adults were removed) corresponding to video fragments.

2.5 Likert Scale

By video, two experts scored for children's performance of test tasks (Cohen kappa coefficient, k = 1.0 [27]) on a 4-point Likert scale [26] "1 = none, 2 = slightly, 3 = moderate, 4 = perfect." The criteria for scoring were: points for emotions recognition: 1 – does not perform the test; 2 – up to 49% correct answers; 3- more than 50% correct answers; 4 - all correct answers. More details are presented in the earlier publication on testing the methodology [25].

2.6 Perceptual Study

Video and audio tests were created for the perceptual experiment. 50 video fragments were selected from the dataset for TD children and a video test was created. For TD children, the selection of fragments of recordings when performing test tasks CEDM, their annotation and number were carried out similarly to the previously used approach for children with ASD, ID [24], and DS [23]. The duration of the fragments was from 3 to 30 s. Before each video fragment, the number was inserted. The pause between the fragments was 10 s. Each video fragment was included in the test once.

The video test was presented to a group of adults (experts) without sound from a personal computer monitor. When watching the test, experts noted the emotional state of the children, choosing one of the four proposed categories.

The audio test contained children's speech corresponding to the video fragments. Each speech signal was repeated once in the test, the pause between speech signals was 5 s. The audio test was presented to experts (the same that watched the video) in an open field. There was no preliminary training of adults. The video and audio tests were used for automatic analysis.

2.7 Automatic Analysis of Facial Expression and Emotional Speech of Children

Analysis of Facial Expression Using Convolutional Neural Network. For preprocessing, the video records were split into series of frames (static images) with FFmpeg [28], a free and open-source software project consisting of a suite of libraries and programs for handling video, audio, and other multimedia files and streams, and then each frame was processed with OpenCV face detector via Deepface [29], a lightweight face recognition and facial attribute analysis framework. Since the video segments with a single child contained only a small number of frames without a detected face, we applied a sliding window to filter out the empty frames, timestamp the segments, and map them to the labels. For the face emotion detection, we used the Deepface default model based on the VGG-face infrastructure.

Analysis of Emotional Speech Using Recurrent Neural Network (RNN). For the audio records, the segments with utterances of a single child were manually clipped via Audacity [30], a free and open-source digital audio editor and mapped to the labels. For the speech emotion detection, we used a simple recurrent neural network with two recurrent and two fully-connected layers with 128 RNN and dense units each trained on a combination of The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), Toronto Emotional Speech Set (TESS), and Berlin Database of Emotional Speech (EMO-DB) datasets. We did not apply any additional transfer learning or fine-tuning.

Confusion matrixes were prepared. We calculated recall, precision, F-1 score for each emotion, Unweighted Average Recall (UAR) - for all emotions [31].

3 Results

3.1 Study 1: Perceptual Experiment

Recognition by Experts of the Emotional State of TD Children by Facial Expression. An analysis of the results of the perceptual experiment showed that by video fragments of TD children, experts recognize the joy (97% correct answer), anger (69%), and neutral (63%) states better than the sadness state, confusing it with the neutral state (45%) (Table 1). The Unweighted Average Recall (UAR) was 0.67.

Table 1. Confusion matrix for recognition by experts of the emotional states of TD children by video fragments (% of expert's answers).

	joy	neutral	sadness	anger
joy	**97**	2	1	0
neutral	7	**63**	25	5
sadness	3	45	**39**	13
anger	6	19	6	**69**
Recall	0.97	0.63	0.39	0.69
Precision	0.86	0.49	0.55	0.79
F1-score	0.91	0.55	0.46	0.74
UAR = 0.67				

Note: The column headers indicate predictions

Table 2. Confusion matrix for recognition by experts of emotional states by speech of TD children (% of expert's answers).

	joy	neutral	sadness	anger
joy	**80**	15	4	1
neutral	10	**82**	7	1
sadness	4	24	**70**	2
anger	4	17	7	**72**
Recall	0.80	0.82	0.70	0.72
Precision	0.82	0.59	0.80	0.95
F1-score	0.81	0.69	0.74	0.82
UAR = 0.76				

Recognition of the Emotional State of TD Children by Speech. The experts recognize all emotions in children's speech with great accuracy, with better recognition of the joy and neutral states (Table 2). The UAR was −0.76. They confused sadness and anger mainly with the neutral state.

3.2 Study 1: Automatic Analysis of Facial Expression

Analysis Using Convolutional Neural Network. By the video of TD children, the neutral and joy states were classified better than the anger and sadness states (Table 3). UAR for TD children was −0.43. The performance is above chance level.

Table 3. Confusion matrix for automatic classification of emotional states of TD children by video fragments, %.

	joy	neutral	sadness	anger
joy	**51**	35	6	8
neutral	3	**66**	23	8
sadness	5	52	**31**	12
anger	3	47	26	**24**
Recall	0.51	0.66	0.31	0.24
Precision	0.76	0.51	0.3	0.38
F1-score	0.61	0.57	0.31	0.29
UAR = 0.43				

3.3 Study 1: Automatic Analysis of Speech

Analysis Using Recurrent Neural Network. For the audio, the joy and neutral states were classified better than the anger and sadness states (Table 4). UAR for TD children was −0.5. The performance for audio is higher than for video.

Table 4. Confusion matrix for automatic classification of emotional states of children by audio, %.

	joy	neutral	sadness	anger
joy	**62**	15	23	0
neutral	8	**54**	23	15
sadness	29	14	**43**	14
anger	10	30	20	**40**
Recall	0.62	0.54	0.43	0.4
Precision	0.57	0.5	0.43	0.5
F1-score	0.59	0.52	0.43	0.44
UAR = 0.50				

3.4 Study 2: Recognition of the Emotional State of TD Children, Children with ASD, ID, DS: Comparative Data

Experts identified the emotions of children from all groups by video and speech more accurately than automatic analysis, with higher UAR values for TD children by audio and video in a perceptual experiment and by audio in the automatic classification of

emotions (Fig. 2). UAR values for children of different groups did not differ significantly in emotion recognition in the video by experts and automatic. Experts identified emotions more accurately in the audio test for TD children vs children with ASD, ID, and DS.

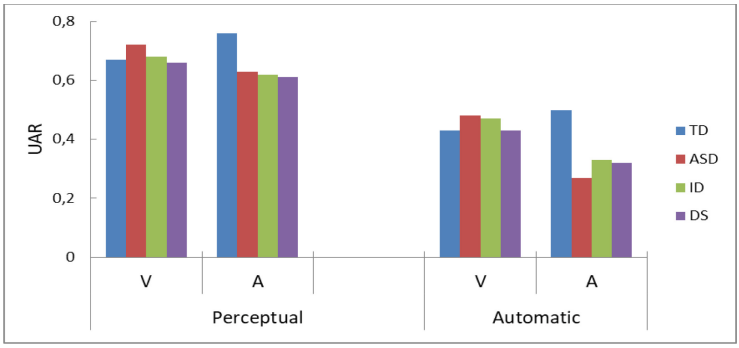


Fig. 2. The Unweighted Average Recall for emotions recognized from video and audio of TD children, children with ASD, ID, DS by experts and automatic.

Experts recognize all emotions from video and audio in children of all groups, but with varying accuracy (Fig. 3 A.C). All emotions in all groups of children are automatically classified by video (Fig. 3 B), with worse recognition of the anger state; by audio - all emotions are classified in TD children; in children with ASD, ID, and DS - only the neutral state is classified with greater accuracy (Fig. 3 D).

Experts better recognized the joy state from the facial expressions of TD children and children with ID, the neutral state in children with ASD and DS, and worse – the state of sadness in all groups of children (Fig. 3 A). Recognition of the anger state in TD children, children with ASD and ID does not differ significantly; experts recognized worse anger from the facial expressions of children with DS.

By the video of ASD, ID and DS children, the joy state was classified automatic better than the neutral, anger and sadness states; the neutral state was classified better in TD children; anger state is poorly classified in all groups of children (Fig. 3 B).

Experts recognized the emotions of children of all groups from audio, with the higher accuracy for all emotions in TD children, for anger - in children with ASD and DS. The joy state is worse recognized in the speech of children with DS (Fig. 3 C).

For the audio of TD children, all emotions were automatic classified with performance is above chance; of ASD, ID and DS children, the neutral state was classified better than the joy, anger, and sadness states. By audio of children with ASD, the joy and sadness states were not classified automatically, the anger state – for children with ID, the sadness state – for children with DS (Fig. 3 D).

Likert Scale Scores for Test Tasks Completed by Children. Children received scores on a Likert scale for completing test tasks in which they naturally (spontaneous) showed different emotions (conversation with the experimenter, playing with toys, co-op play, telling a story based on pictures "Tale of the little Lion") and for "acting" expressions of emotions. "Acting" play is the test task in which a child is asked to express the emotions

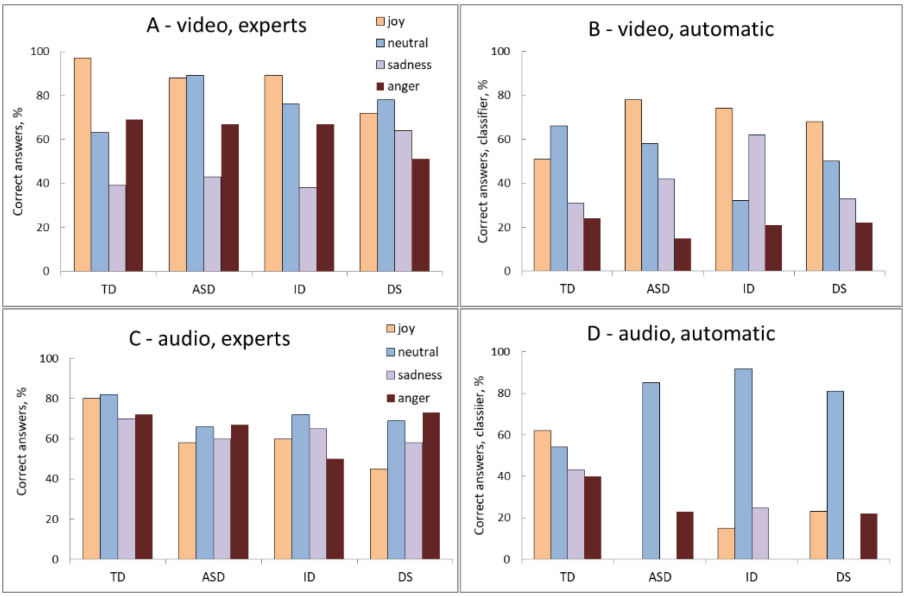


Fig. 3. Recognition of the emotional states of TD children, children with ASD, ID and DS by their facial expressions and speech by experts and automatic, %.

"joy – neutral – sadness – anger" in the voice and in the facial expression. In test tasks with spontaneous expression of emotions, children with ASD had the lowest scores compared to children with DS, ID and, especially, TD, for the expression of emotions (1–4 points) and for a fully completed test task (Fig. 4 A).

Emotions manifestation by the voice in "acting" play was a more difficult test task for children of all groups compared to manifesting emotions in facial expression, with higher scores in TD children (Fig. 4 B). Children with DS had the lowest scores for the test task – emotions expression in voice and facial expressions; children with ASD had lower scores than children with ID and TD.

3.5 Emotional Portrait of Children: Summary

For TD Children. Experts recognized all emotions by video and audio with great accuracy, sadness - worse by video with less accuracy in automatic classification of sadness and anger states. On the Likert scale, high scores for completed test tasks, spontaneous emotions, emotions in facial expressions and voice when performing test tasks in "acting" play were obtained.

For Children with ASD. Experts recognized by video joy and a neutral states with great accuracy, anger - worse, sadness – poorly. The joy state was classified automatically with greater accuracy, neutral state and sadness - worse, with the worst recognition for the anger state. By audio, experts recognized all emotions well, with more accurate recognition of the anger state; in automatic classification, the neutral state was better classified, anger was poorly classified, and joy and sadness were not recognized.

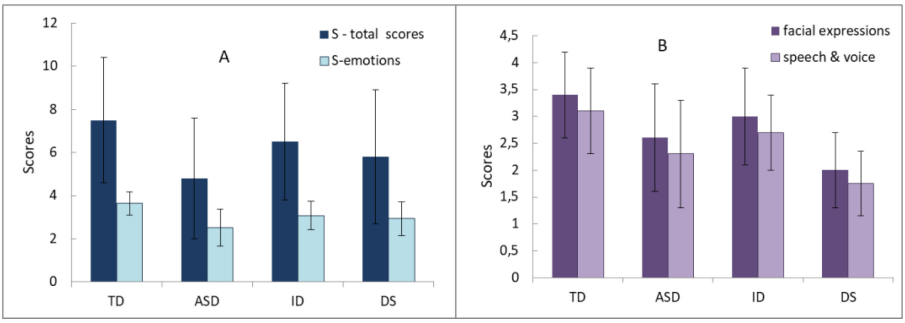


Fig. 4. Likert scale scores for: A - scores for test tasks in which children naturally (spontaneously) expressed different emotions. Dark columns - total points for the test task, light columns - points for reflection different emotions during performing the test task. B - points for "acting" play – emotions manifestation in facial expressions (dark columns) and in speech & voice (light columns).

On the Likert scale, the lowest scores are for completed test tasks and spontaneous emotions; in "acting" play test tasks, children express emotions in their voice and facial expressions (worse than in their voice), but better than children with DS and worse than children with ID.

For Children with ID. Experts better recognized by video joy, neutral, anger states, worse – sadness state. All emotions were automatically classified by video, with the highest accuracy for joy and sadness states, worse – for neutral, and worst – for anger state. By audio, experts recognized all emotions, with worse recognition of anger. The neutral state was automatically classified, joy and sadness were classified worse, and anger was not classified.

On the Likert scale, scores for completed test tasks and spontaneous emotions were higher than for children with DS and ASD, lower than those for TD children. In "acting" play test tasks, children showed all emotions well in facial expressions and voice; scores were higher than those of ASD children and children with DS.

For Children with DS. By video, experts recognized joy, neutral, and sadness states, with worse recognition of the anger state; automatic system classified all emotions, better - joy and sadness, worse - neutral, and worst - anger state. By audio, experts recognized all emotions, with the worst recognition of joy; automatic system classified neutral state with great accuracy, worse – joy and anger, does not recognize sadness.

On the Likert scale, scores for test tasks and spontaneous emotions were higher than for children with ASD, lower than for children with ID. In test tasks for "acting" play, children with DS showed emotions poorly in facial expressions and voice; their scores were lower than those for ASD and ID children. Scores for test tasks and spontaneous emotions were higher, in test tasks for "acting" play there were low scores for the manifestation of emotions in facial expressions and voice.

4 Discussion and Conclusion

The present work is part of the investigation on the formation of emotions in children with typical development and children with developmental disorders. Children's ability to express emotions in facial expressions, voice and speech was estimated through human assessment, automatic classification, and Likert scale scores. The paper presents the results of two experimental studies – the recognition of emotional states of TD children by humans and automatically, comparative analysis of the recognition of the emotional states of children with different groups (TD, ASD, ID, DS) by video and speech. For the correct comparison of the results, a unified approach to the collection and analysis of material [25] was used.

Different approaches to the analysis of the recognition of the emotions in children - a perceptual experiment, automatic classification by facial expression and speech, and Likert scale scores revealed that the emotions of TD children are recognized better vs ID, ASD, and DS children. This result was expected, but it is important as a control with which data for children with ASD, ID, and DS can be compared. Experts and automatic systems classified emotions more accurately by the audio in TD children compared to children with ASD, ID, and DS. Sadness and anger states in TD children were recognized less accurately than joy and neutral states, which may be due to cultural specificity [32] and upbringing [33].

The question of the correctness of comparing the manifestations of emotions in children with ASD and TD children is discussed [34, 35]. It was shown that the facial expressions of children with ASD are considered unusual, which is represented in the questionnaires used for autism diagnosing [34]. The emotions of children with ASD are more diverse than the emotions proposed in the questionnaire of the perceptual experiment and are more reminiscent of the complex emotional manifestations of adults [35]. Therefore, in this study, we use data on TD children as a baseline, and compare the expression of emotions in children of different groups - ASD, ID, and DS.

We assumed that the low accuracy of recognition of any emotion in one modality could be compensated by more accurate recognition in another modality. Thus, for each emotional state of the child of each developmental disorder - ASD, ID, and DS, the correct classification ranges were obtained. Likert scale scores for completed test tasks would be used to further assess the development of emotions in children. Additionally, since scores are assigned for each test task completed, they can more accurately indicate which task the child is performing worse or better. The final result would be a judgment, based on taking into account the range of accuracy of emotion classification and scoring, which category the analyzed emotional manifestations belong to – closer to TD, closer to ASD, closer to ID or DS. But the results showed the following:

For children with ASD the neutral state was automatically classified with great accuracy by video and audio, poorly – anger state. Anger, which is poorly recognized by video, can be supplemented by recognition by audio. Joy and sadness states are well classified only by video.

Automatic recognition of emotions in children with ID from video can be supplemented by recognition from the audio modality, since from video - the neutral state is poorly recognized, but from audio – good. The joy and sadness are good recognized from video and poorly from audio; anger is recognized only by video.

For children with DS, the joy state was automatically classified with great accuracy by video (correspond with data [21]), neutral state – by video and audio. The states of sadness and, especially, anger are poorly classified by video; the state of sadness is not classified by audio. The results of automatic classification also correspond to scores on a Likert scale.

Thus, emotions with a negative valence - sadness and anger - are automatically classified poorly by audio and video in children with ASD, ID, and DS, which is consistent with the works about difficulties in classifying emotions in individuals with ASD [36] and ID [37]. Based on audio, children's neutral state is classified with great accuracy.

At least two questions arise. First, why is anger automatically classified poorly, but recognized well by people? For European culture, the expression of anger is socially unacceptable; children are taught not to show negative emotions [6]. Automatic classification methods have shown that negative emotions among representatives of different cultures are well recognized [38]. People, knowing the nuances of the manifestation of anger, rely on weakly expressed characteristics of facial expressions or voices to recognize anger. Second, what could be the reason for this - whether it is due to the approaches used for automatic classification or the material, especially for the video modality? Previously, we raised the question of using photographs to recognize emotions by facial expression [35], but such an approach would significantly simplify the task. It is known that the movement of facial features provides additional unique temporal information that contributes to more accurate emotion recognition [39]. This poses the task of finding new algorithms for classifying the emotions of children with developmental disabilities based on their facial expressions and speech. The novelty of the results lays in the obtaining normative data on the recognition of emotions in TD children and in comparative data on groups of TD children, children with ASD, ID, DS.

Limitation: 1. Small number of audio and video fragments for automatic classification. 2. We used existing software to classify the emotional states of TD children and children with DS, ASD, ID for the purpose to compare the data.

Future: 1. Multimodal classification models (Audio (voice and speech) + Video) 2. Creation new approaches for automatic classification of emotional state of children with atypical development, taking into account their psychoneurological state.

Acknowledgements. This study is financially supported by the Russian Science Foundation (project 22–45-02007) - for Russian researches, DST/INT/RUS/RSF/P-57/2021 – for Indian researches.

References

1. Bargagna, S., et al.: Educational robotics in down syndrome: a feasibility study. Technol. Knowl. Learn. **24**, 315–323 (2019)
2. Alemi, M., Bahramipour, S.: An innovative approach of incorporating a humanoid robot into teaching EFL learners with intellectual disabilities. Asian-Pac. J. Second Foreign Lang. Educ. **4**, 10 (2019)

3. Schadenberg, B.R., Reidsma, D., Heylen, D.K.J., Evers, V.: Differences in spontaneous interactions of autistic children in an interaction with an adult and humanoid robot. Front. Robot. AI **7**(28), 1–19 (2020)
4. Garg, R., et al.: The last decade of HCI research on children and voice-based conversational agents. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22), Article 149, pp. 1–19. New York, NY, USA (2022)
5. Tsai, Y.T., Lin, W.A.: Design of an intelligent cognition assistant for people with cognitive impairment. In: IEEE 20th International Conference on High Performance Computing and Communications, 16th International Conference on Smart City and 4th International Conference on Data Science and Systems, HPCC/SmartCity/DSS 2018, pp.1207–1212. IEEE, Exeter, UK (2018)
6. Craig, G.J., Baucum, D.: Human Development. 9th edn. Pearson College Div, US (2001)
7. Ma, W., Zhou, P., Liang, X., Thompson, W.F.: Children across cultures respond emotionally to the acoustic environment. Cogn. Emot. **37**(6), 1144–1152 (2023)
8. Lyakso, E., et al.: Recognition of the emotional state of children by video and audio modalities by Indian and Russian experts. LNAI **14338**, 469–482 (2023)
9. Surian, D., van den Boomen, C.: The age bias in labeling facial expressions in children: effects of intensity and expression. PLoS ONE **17**(12), e0278483 (2022)
10. Fridenson-Hayo, S., et al.: Basic and complex emotion recognition in children with autism: cross-cultural findings. Mol. Autism **7**, 52 (2016)
11. Russel, J.A., Bachorowski, Jo-A., Fernandez-Dols, J-M.: Facial and vocal expressions of emotion. Annu. Rev. Psychol. **54**(1), 329–349 (2002)
12. Reyes, N.M., Pickard, K., Reaven, J.: Emotion regulation: a treatment target for autism spectrum disorder. Bull. Menninger Clin. **83**(3), 205–234 (2019)
13. Reyes, N.M., Factor, R., Scarpa, A.: Emotion regulation, emotionality, and expression of emotions: a link between social skills, behavior, and emotion problems in children with ASD and their peers. Res. Dev. Disabil. **106**, 103770 (2020)
14. Philip, R.C., et al.: Deficits in facial, body movement and vocal emotional processing in autism spectrum disorders. Psychol. Med. **40**(11), 1919–1929 (2010)
15. Yeung, M.K.: A systematic review and meta-analysis of facial emotion recognition in autism spectrum disorder: the specificity of deficits and the role of task characteristics. Neurosci. Biobehav. Rev. **133**, 104518 (2022)
16. Frolova, O., Nikolaev, A., Grave, P., Lyakso, E.: Speech features of children with mild intellectual disabilities. In: Companion Publication of the 2023 International Conference on Multimodal Interaction (ICMI '23 Companion), pp. 406–413. ACM, New York, NY, USA (2023)
17. Goldbart, J.: Communication as a human right for children with profound intellectual disabilities. Dev. Med. Child Neurol. **65**(6), 725–726 (2023)
18. Roch, M., Pesciarelli, F., Leo, I.: How individuals with down syndrome process faces and words conveying emotions? Evidence from a priming paradigm. Front. Psychol. **11**, 692 (2020)
19. Pochon, R., Declercq, C.: Emotion recognition by children with down syndrome: a longitudinal study. J. Intellect. Dev. Disabil. **38**(4), 332–343 (2013)
20. Lyakso, E., Frolova, O., Gorodniy, V., Grigovev, A., Nikolaev, A., Matveev, Y.: Reflection of the emotional state in the characteristics of voice and speech of children with Down syndrome. In: Proceedings SpeD 2019, 10th IEEE International Conference on Speech Technology and Human-Computer Dialogue, pp. 1–6. Timisoara, Romania (2019)
21. Carvajal, F., Iglesias, J.: Judgements of facial and vocal signs of emotion in infants with down syndrome. Dev. Psychobiol. **48**(8), 644–652 (2006)
22. Dykens, E., Hodapp, R.M., Evans, D.W.: Profiles and development of adaptive behavior in children with down syndrome. Am. J. Ment. Retard. **98**(5), 580–587 (1994)

23. Lyakso, E., et al.: Recognition of the emotional state of children with down syndrome by video, audio and text modalities: human and automatic. LNAI **13721**, 438–450 (2022)
24. Lyakso, E., et al.: Emotional state of children with ASD and intellectual disabilities: Perceptual experiment and automatic recognition by video, audio and text modalities. LNAI **14338**, 535–549 (2023)
25. Lyakso, E., Frolova, O., Kleshnev, E., Ruban, N., Mekala, M., Arulalan, K.V.: Approbation of the child's emotional development method (CEDM). In: Companion Publication of the 2022 International Conference on Multimodal Interaction (ICMI '22 Companion), pp. 201–210. New York, NY, USA (2022)
26. Likert, R.: A technique for the measurement of attitudes. Arch. Psychol. **22**, 5–55 (1932)
27. Md Juremi, N.R., Zulkifley, M.A., Hussain, A., Zaki W.M.D.: Inter-rater reliability of actual tagged emotion categories validation using Cohen's Kappa coefficient. J. Theor. Appl. Inf. Technol. **95**, 259–264 (2017)
28. FFmpeg. https://ffmpeg.org. Accessed 02 Jul 2024
29. Multi-task Cascaded Convolutional Networks (MTCNN) via Deepface. https://github.com/serengil/deepface. Accessed 02 Jul 2024
30. Audacity. https://www.audacityteam.org. Accessed 02 Jul 2024
31. Dalianis, H.: Evaluation metrics and evaluation. In: Clinical Text Mining, pp. 45–53. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-78503-5_6
32. Lyakso, E., Ruban, N., Frolova, O., Mekala, M.A.: The children's emotional speech recognition by adults: cross-cultural study on Russian and Tamil language. PLoS ONE **18**(2), e0272837 (2023)
33. Kaya, H., Salah, A.A., Karpov, A., Frolova, O., Grigorev, A., Lyakso, E.: Emotion, age, and gender classification in children's speech by humans and machines. Comput. Speech Lang. **46**, 268–283 (2017)
34. Jacques, C., Courchesne, V., Mineau, S., Dawson, M., Mottron, L.: Positive, negative, neutral- or unknown? The perceived valence of emotions expressed by young autistic children in a novel context suited to autism. Autism **26**(7), 1833–1848 (2022)
35. Lyakso, E.E., Frolova, O.V., Grigorev, A.S., Sokolova, V.D., Yarotskaya, K.A.: Recognition by adults of emotional state in typically developing children and children with autism spectrum disorders. Neurosci. Behav. Physiol. **47**(9), 1051–1059 (2017)
36. Landowska, A., et al.: Automatic emotion recognition in children with autism: a systematic literature review. Sensors (Basel) **22**(4), 1649 (2022)
37. Hammann, T., et al.: The challenge of emotions — an experimental approach to assess the emotional competence of people with intellectual disabilities. Disabilities **2**, 611–625 (2022)
38. Hughson, E., Javadi, R., Thompson, J., Lim, A.: Investigating the role of culture on negative emotion expressions in the wild. Front. Integr. Neurosci. **15**, 699667 (2021)
39. Ambadar, Z., Schooler, J.W., Cohn, J.F.: Deciphering the enigmatic face: the importance of facial dynamics in interpreting subtle facial expressions. Psychol. Sci. **16**(5), 403–410 (2005)

Speech Processing for Medicine

Investigating the Utility of wav2vec 2.0 Hidden Layers for Detecting Multiple Sclerosis

Gábor Gosztolya[1,2(✉)], László Tóth[2], Veronika Svindt[3], Judit Bóna[4], and Ildikó Hoffmann[3,5]

[1] HUN-REN–SZTE Research Group on Artificial Intelligence, Szeged, Hungary
ggabor@inf.u-szeged.hu
[2] Institute of Informatics, University of Szeged, Szeged, Hungary
[3] HUN-REN Research Center for Linguistics, Budapest, Hungary
[4] Department of Linguistics, ELTE Eötvös Loránd University, Budapest, Hungary
[5] Department of Psychiatry, University of Szeged, Szeged, Hungary

Abstract. Multiple sclerosis (MS) is a chronic autoimmune neurodegenerative disease, affecting the central nervous system. The disease can induce various symptoms, such as adversarily affecting the speech of the subject in various ways, therefore allowing the use of automatic speech analysis for the detection of MS and for monitoring the condition of the patient. Owing to data scarcity, however, deep neural networks are usually not employed for this task as classifiers, but are used as feature extractors. This is the case for self-supervised networks such as wav2vec 2.0 as well, where a straightforward source of embeddings (used as features) are the last layers of the convolutional (lower) and fine-tuned (higher) blocks. In this study we investigate whether extracting the embeddings from some other, inner layer of the fine-tuned (transformer) block can help improve MS detection performance. Tested on two speech tasks, we found that the lowest one-third of the 24 fine-tuned layers proved to be the most suitable for feature extraction, which led to statistically significant improvements in the AUC scores for both speech tasks.

Keywords: Multiple sclerosis · Pathological speech processing · Wav2vec 2.0 · Feature extraction

1 Introduction

Multiple sclerosis (MS) is a chronic autoimmune neurodegenerative disease, affecting the central nervous system, which can result in various cognitive and linguistic impairments of the subjects [29]. The progression of MS may vary considerably from subject to subject, and it can change over time. Several changes may occur as the disease progresses: an increase in disability (affecting walking,

A. Karpov and V. Delić (Eds.): SPECOM 2024, LNAI 15299, pp. 297–308, 2024.
https://doi.org/10.1007/978-3-031-77961-9_22

balance, coordination, and other physical abilities of the patient); an increase in fatigue; sensory changes (affecting the ability to feel cold, heat and touch) and changes in cognitive and language functions. Noting these points, automatic speech analysis might contribute to detect the disease in an automatic, contact-free and (relatively) cheap way, or serve as a screening technique.

In the past decade, automatic speech analysis has developed into a broad area within speech technology. It includes *computational paralinguistics*, which seeks to automatically identify different speaker traits and states, such as emotion recognition [13,21], speaker age and gender determination [22], assessing the degree of sleepiness [15], whether the speaker has cold [33], or the presence of stuttering [12]. It also includes *pathological speech processing* tasks, where the aim is to automatically decide whether the speaker is suffering from a specific disease such as Parkinson's Disease [17,19], Alzheimer's Disease [16,27], mild cognitive impairment [25] or depression [9,18]. After the deep learning revolution, deep networks also found their way into the pathological speech processing area [9, 17,27].

Nowadays, with the emergence of self-supervised learning, perhaps the most widely-used speech processing network type is wav2vec 2.0 [2]. Besides direct speech processing applications [6,24], evaluating the network on a specific speech utterance and noting the activations of a specific hidden layer (i.e. the *embeddings*) and using these vectors as features (and thus, the whole network as a feature extractor) is a common approach as well [7,26]. Due to the scarcity of resources in the pathological speech processing area, deep networks are rarely used as classifiers there, but they primarily serve as feature extractors [7,27,30].

To employ neural networks (including those with a wav2vec 2.0 architecture) as feature extractors, one has to choose a specific layer to take the embeddings from. A wav2vec 2.0 network has two main blocks: the lower *convolutional* one and the higher *fine-tuned* one, and the straightforward sources of embeddings are the last layers of each block [20]. In this study, however, we investigate whether some inner layer of the fine-tuned block might supply better features. For this, we take a network fine-tuned on the target language (in our case, Hungarian), and test the embeddings taken from all of the inner layers of the fine-tuned block as machine learning features to distinguish multiple sclerosis patients from healthy control (HC) subjects.

2 The Hungarian Multiple Sclerosis Corpus

All the tests were carried out at the Neurology Department of Uzsoki Hospital, Budapest, Hungary, and at the Research Center for Linguistics of the Hungarian Research Network, Budapest, Hungary. The study was approved by the Ethics Committee of the Uzsoki Hospital, and it was conducted in accordance with the Declaration of Helsinki. In the current study we use the recordings of 23 MS subjects (5 males and 18 females) and 22 healthy controls (6 males and 16 females). All 23 MS subjects belonged to the relapsed-remitting MS subtype (RRMS). All the speakers involved in the study were native Hungarian speakers. The MS and

HC groups displayed no statistically significant difference in their demographic attributes (age in years, gender (male / female) and years of education).

The protocol for collecting the speech samples from the subjects was quite extensive, involving 17 different speech tasks. In the current study, due to space limitations, we use the recordings of two spontaneous speech tasks: in the **Opinion** task the subjects were asked to share their opinions about vegetarianism, while in the **Narrative Recall** speech task the subjects listened to a two-minute-long historical anecdote that was unknown to them beforehand, and they had to summarize the story heard as accurately as possible. Although participants have to produce coherent, complex narratives in both tasks, there are some significant differences in the cognitive requirements of these task. Namely, in the Narrative Recall task the speakers had to rely significantly on their working memory, and they had to inhibit irrelevant information, compared to the clearly simpler Opinion speech task.

The recording was performed with a Sony PCM-A10 digital dictaphone using a tie clip microphone with a sampling rate of 48 kHz; later the recordings were converted to 16 kHz mono with a 16 bit resolution.

3 Wav2vec 2.0

wav2vec is a convolutional neural network (CNN) designed to process raw audio signals as input and generate representations suitable for automatic speech recognition (ASR) systems. The model is trained in a self-supervised manner, during which it learns to predict future observations for the given speech sample [28]. This self-supervised training allows the model to be pre-trained on large, unannotated corpora, enabling subsequent fine-tuning for specific audio processing tasks such as ASR for low-resource languages [24] or paralinguistic appications (e.g. emotion detection [26]). The **wav2vec 2.0** architecture further enhances this approach by incorporating masking during training. Specifically, raw audio is encoded using a block of convolutional neural networks, and small segments of the resulting latent speech representations are masked, akin to masked language modeling. These masked representations are then processed by a quantizer, which selects speech units from an inventory of learned units, and a transformer network, which incorporates information from the entire utterance [2]. Figure 1 shows the layout of the (fine-tuned) wav2vec 2.0 structure.

3.1 Wav2vec2 for Feature Extraction

The outputs from the multi-layer convolutional block are the sequence of extracted feature vectors of the last convolutional layer, while the outputs from the second (fine-tuned) block comprise the sequence of the hidden states of the last layer of the block. These two types of feature vectors may carry relevant information for a large range of speech processing tasks, so they are quite popular as features [8,20]. Of course, these embeddings are at the frame level, so the number of these vectors is proportional to the length of the utterance. To

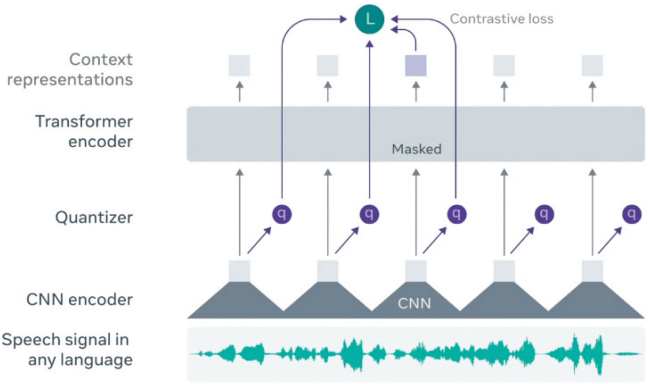


Fig. 1. The fine-tuned wav2vec 2.0 framework structure. Source: https://ai.meta.com/blog/wav2vec-20-learning-the-structure-of-speech-from-raw-audio.

employ them as *utterance-level* features, they have to be aggregated over the whole recording. To do this, taking the mean and/or the standard deviation of the values over the whole utterance is a generally accepted solution [10,27,32].

In this study we, however, focus on the *inner* layers of the fine-tuned block. Since in a standard XLSR-53 network (see e.g. [1]), there are 24 such transformer layers, we have 24 options of feature extraction. Besides the standard assumptions that lower-laying layers capture lower-level phenomena (e.g. silence, noise, acoustic conditions), while higher-level layers tend to capture high-level (e.g. phonetic) information, we do not have any further guidance. Due to this, in our experiments we tested the activations taken from all 24 layers for the two speech tasks for multiple sclerosis detection.

4 Experimental Setup

4.1 Feature Extraction

We used the wav2vec 2.0 model `wav2vec2-large-xlsr53-hungarian`. The base of this model is the `XLSR-53` model pre-trained by Facebook on the audio data of 53 languages simultaneously, and it was ensured that the quantization module of the wav2vec 2.0 neural network also delivers multilingual quantized speech units [1]. This base model was then fine-tuned by the user `jonatasgrosman` [11] on the Hungarian part of the Mozilla Common Voice 6.1 corpus (8 h). The last layer of the convolutional block of this model consists of 512 neurons, while all the layers of the fine-tuned block have 1024 neurons. By using mean and standard deviation of these frame-level embedding vectors, we obtained 1024 and 2048 utterance-level features, convolutional and fine-tuned embeddings, respectively. Since we focused on the fine-tuned embeddings, in most of our experiments we had feature vectors with a length of 2048.

4.2 Utterance-Level Classification

We employed the approach common in pathological speech processing studies (e.g. [3,14,31]): due to the low number of examples (subjects) from a machine learning perspective, we did not define separate training, development and test sets, but used cross-validation. Each fold consisted of the data of one MS and one HC speaker, leading to 23 folds overall. Classification performance was measured using the Area Under the ROC Curve (AUC) metric, also commonly applied in pathological speech processing studies [3,10].

We employed Support Vector Machines (SVM) for classification, using the LibSVM [5] library. We employed the nu-SVM method with a linear kernel; the value of C was tested in the range $10^{\{-5,\ldots,1\}}$. The optimal value for the C meta-parameter was determined by the technique called *nested cross-validation* [4]: for the speakers of 22 folds in the training subset of the actual CV step, we performed *another* cross-validation. We chose the C value that gave the highest AUC score in this "inner" cross-validation loop; the "final" SVM model was then trained on the data of 22 folds with this C value, and it was evaluated on the data of the last fold (i.e. two speakers). With this procedure we sought to avoid the bias in our scores that would have been present if we had used standard cross-validation.

To measure the robustness of the AUC scores, we repeated each classification experiment five times, using a different random seed value when assigning the speakers to specific folds for cross-validation. In the results, we report the mean of the five AUC scores. When inspecting robustness, we calculate the standard deviation (*Std.*) of the five AUC values, and report the range (i.e. $[\min, \max]$) of the scores as well.

5 Results with the Embeddings of the Last Layers

Table 1 shows the results obtained for both speech tasks when we used the embeddings from the last layers of the convolutional and the fine-tuned blocks. In general, the results are acceptable, with AUC values lying between 0.654 and 0.824, and mean AUC scores between 0.707 and 0.806. Focusing on the mean

Table 1. AUC values obtained for the two speech tasks, when using the embeddings from the last layers of the convolutional and the fine-tuned blocks. AUC is reported as the average (*Mean*) of the five values measured with the five random speaker fold assignments, along with the standard deviation (*Std.*) and the range ([min, max]).

Speech task	Embedding type	AUC		
		Mean	Std.	Range
Opinion	Convolutional	0.707	0.032	[0.654, 0.737]
	Fine-tuned	0.736	0.025	[0.698, 0.763]
Narrative Recall	Convolutional	0.724	0.008	[0.712, 0.733]
	Fine-tuned	0.806	0.014	[0.787, 0.824]

AUC values (averaged over the five classification runs, constructing the folds from different speaker pairs), we can see that the embeddings taken from the last layer of the fine-tuned block outperform those from the convolutional block. Furthermore, the *Opinion* speech task was less effective for detecting multiple sclerosis than the *Narrative Recall* task, since the mean AUC values were higher for both embedding types. When inspecting the standard deviations and the ranges of the AUC scores, we also see that the variance was definitely smaller for the *Narrative Recall* speech task than for the *Opinion* task, suggesting a more robust classification performance. Of course, as the main focus of our investigation is the embeddings taken from the inner fine-tuned layers, the values presented in Table 1 serve only as reference values.

6 Results with the Embeddings of the Inner Layers

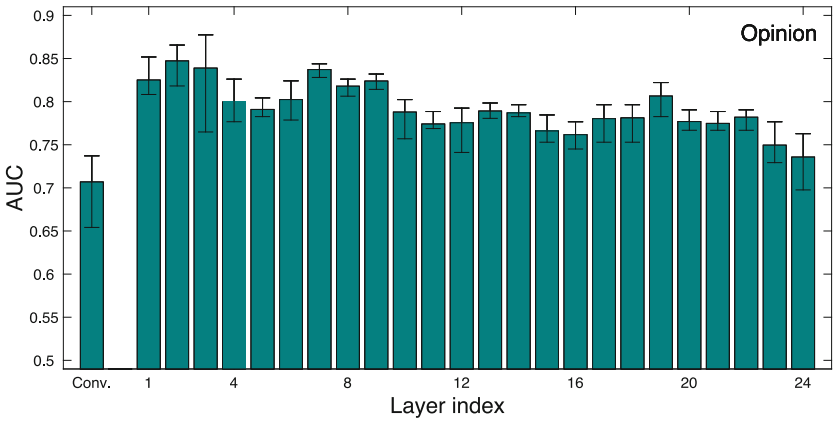


Fig. 2. Mean AUC values (bars) and the [min, max] range (error bars) obtained when using the embeddings from the last layer of the convolutional block (*Conv.*), and when using the hidden layers of the fine-tuned block, for the Opinion speech tasks.

Figure 2 shows the mean AUC values obtained for the *Opinion* speech task for the last layer of the convolutional block (*Conv*) and for all the layers of the fine-tuned block (1...24). (Of course, the 24^{th} layer is the last layer of the fine-tuned block, i.e. that shown in Table 1) Quite surprisingly, the embeddings taken from *any* inner layers (i.e. 1...23) outperformed both those taken from the convolutional layer and those taken from the last fine-tuned layer. The difference, measured by the Mann-Whitney U test (see [23], also known as the Wilcoxon rank-sum test), was statistically significant in almost all cases, with only three exceptions (the embeddings of the 15^{th}, 16^{th} and the 23^{th} layers relative to the last layer of the fine-tuned block). In general, lower layers tend to work better

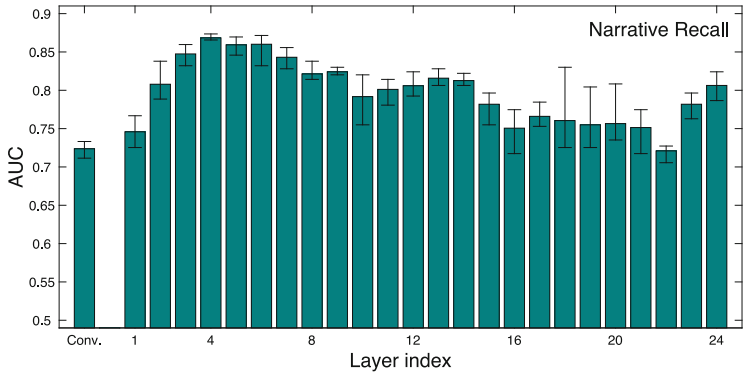


Fig. 3. Mean AUC values (bars) and the [min, max] range (error bars) obtained when using the embeddings from the last layer of the convolutional block (*Conv.*), and when using the hidden layers of the fine-tuned block, for the Narrative Recall speech task.

than those in the higher regions of the fine-tuned block, and we measured the highest mean AUC score with the 2^{nd} layer.

We observe similar tendencies for the *Narrative Recall* task (see Fig. 3), although here the last layer of the fine-tuned block was more competitive. The inner layers outperformed the convolutional embeddings statistically significantly in 20 cases (with the exception of the 16^{th}, 18^{th} and 22^{nd} layers), but, compared to the last fine-tuned layer, the improvement (if any) was statistically significant only in 6 cases (lying in the $1 \ldots 9$ region). This suggests that it is worth exploring the inner hidden layers of the fine-tuned block for feature extraction, and that the lower hidden layers of this block might be more useful than those higher up in the wav2vec 2.0 structure, at least for detecting multiple sclerosis.

Table 2 shows the AUC values measured for some specific inner layers of the fine-tuned block; * and ** indicate a significant difference, $p < 0.05$ and $p < 0.01$, respectively, while "—" means there is no statistically significant improvement compared to the reference values. Symbols before and after the slash symbol (i.e. "/") show the difference compared to the last layer of the convolutional and the fine-tuned block, respectively. For the *Opinion* speech task, we obtained the best results with the 2^{nd} hidden layer (mean AUC value of 0.847), while this was the 4^{th} layer for the *Narrative Recall* task. Both variations brought significant improvements over both reference values (with $p < 0.01$), with absolute improvements of 0.111 and 0.060, Opinion and Narrative Recall speech tasks, respectively. Inspecting the standard deviation and range values, we can also see that classification models relying on the embeddings of these inner layers as features are somewhat more robust, having smaller standard deviation scores. In particular, for the Narrative Recall speech task, when we used the embeddings from the 4^{th} hidden layer, the five AUC values fell into a narrow range (0.866, 0.874).

Table 2. Area Under the ROC Curve (AUC) values obtained for the two speech tasks, when using the embeddings from specific fine-tuned layers. Here, * and ** indicate a statistically significant difference ($p < 0.05$ and $p < 0.01$, respectively), while "—" indicates there is no such difference.

Speech task	Embedding type	AUC		
		Mean	Std.	Range
Opinion	Fine-tuned (#2)$^{**/**}$	0.847	0.018	[0.818, 0.866]
	Fine-tuned (#4)$^{**/**}$	0.800	0.023	[0.777, 0.826]
	Fine-tuned (#6)$^{**/**}$	0.802	0.019	[0.779, 0.824]
	Fine-tuned (#8)$^{**/**}$	0.818	0.008	[0.806, 0.826]
	Last convolutional	0.707	0.032	[0.654, 0.737]
	Last fine-tuned	0.736	0.025	[0.698, 0.763]
Narrative Recall	Fine-tuned (#2)$^{**/—}$	0.808	0.022	[0.789, 0.838]
	Fine-tuned (#4)$^{**/**}$	0.868	0.004	[0.866, 0.874]
	Fine-tuned (#6)$^{**/**}$	0.860	0.016	[0.832, 0.872]
	Fine-tuned (#8)$^{**/—}$	0.821	0.010	[0.814, 0.838]
	Last convolutional	0.724	0.008	[0.712, 0.733]
	Last fine-tuned	0.806	0.014	[0.787, 0.824]

Lastly, Table 3 shows the mean, standard deviation and range of all the AUC values obtained for the lower, middle and top one-third of the fine-tuned layers. (In this table the range property is calculated by taking the 5^{th} and 95^{th} percentiles of the 40 AUC scores.) We can say that all three blocks significantly outperformed the convolutional layer, which is not surprising—as it appears that the convolutional embeddings are just too low-level to serve as a base for effective multiple sclerosis detection. Regarding the comparison with the last layer of the fine-tuned block, however, the lowest region of the fine-tuned block is the only one that is sufficiently robust. Although for the Opinion speech task, all three regions gave an improvement with a $p < 0.01$ significance level, for the Narrative Recall speech task only the layers #1...#8 brought a significant improvement ($p = 0.0377$). The middle region (layers #9...#16) were just on par with the last hidden layer, while the topmost one-third of the fine-tuned layers actually led to a significant decrease in the AUC values. This, in our opinion, suggests that the embeddings from the lower layers are, in general, more suited for automatic multiple sclerosis detection, but they still have to come from the fine-tuned block, as embeddings from the convolutional block performed the worst of all configurations tested for both speech tasks.

Table 3. Area Under the ROC Curve (AUC) values obtained for the two speech tasks, when using the embeddings from specific regions of fine-tuned layers. Here, * and ** indicate a statistically significant difference ($p < 0.05$ and $p < 0.01$, respectively), while "—" indicates there is no such difference.

Speech task	Embedding type	AUC		
		Mean	Std.	Range
Opinion	Fine-tuned (#1...#8)$^{**/**}$	0.820	0.028	[0.778, 0.867]
	Fine-tuned (#9...#16)$^{**/**}$	0.783	0.022	[0.749, 0.827]
	Fine-tuned (#17...#24)$^{**/**}$	0.773	0.025	[0.729, 0.810]
Narrative Recall	Fine-tuned (#1...#8)$^{**/*}$	0.832	0.040	[0.740, 0.872]
	Fine-tuned (#9...#16)$^{**/—}$	0.798	0.026	[0.741, 0.828]
	Fine-tuned (#17...#24)$^{**/—}$	0.762	0.033	[0.719, 0.817]

7 Conclusion and Discussion

In this study we investigated whether multiple sclerosis could be automatically detected from the speech of the subjects. For this, we built a workflow consisting of a wav2vec 2.0 model for feature extraction and an SVM model for classification. We used the speech recordings of 45 native Hungarian speakers (23 MS patients of the relapsing-remitting subtype, and 22 healthy controls), performing two spontaneous speech tasks. Besides using the last layers of the convolutional and the fine-tuned blocks of the wav2vec 2.0 model, we experimented with the other hidden layers of the fine-tuned block as potential sources of the embedding vectors. We found that most inner layers were indeed more effective than the final layers of the two blocks: we achieved statistically significant improvements over the convolutional embeddings in 43 cases out of 46, while the last fine-tuned layer was significantly outperformed in roughly half the cases (i.e. 26 times out of 46). Regarding tendencies, we found that the lower-lying hidden layers were more effective for both speech tasks, indicating that lower-level information might be more suitable for multiple sclerosis detection than high-level one, but the convolutional layers alone cannot capture this information. The reason for this might lie in the efficiency of transformers, which are present only in the fine-tuned block.

Of course, the information stored by the independent layers is not mutually exclusive. Therefore, it might be worth using the embedding vectors obtained from different hidden layers in some way, as this combination might improve the classification performance further. However, a fair validation of such combination algorithms might require more subjects than our 45 (which, in other respects, is a fair number of speakers in the pathological speech processing area). Still, we aim to perform such combination experiments in the near future.

Acknowledgments. This study was supported by the NRDI Office of the Hungarian Ministry of Innovation and Technology (grants K-132460 and TKP2021-NVA-09), and within the framework of the Artificial Intelligence National Laboratory Program (RRF-2.3.1-21-2022-00004).

The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Babu, A., et al.: XLS-R: Self-supervised cross-lingual speech representation learning at scale. In: Proceedings of Interspeech, pp. 2278–2282 (2022)
2. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: a framework for self-supervised learning of speech representations. Adv. Neural Inf. Process. Syst. **33**, 12449–12460 (2020)
3. Carvajal-Castaño, H.A., Pérez-Toro, P.A., Orozco-Arroyave, J.R.: Classification of Parkinson's Disease patients - a deep learning strategy. Electronics **11**(17), 2684 (2022). https://doi.org/10.3390/electronics11172684
4. Cawley, G.C., Talbot, N.L.C.: On over-fitting in model selection and subsequent selection bias in performance evaluation. J. Mach. Learn. Res. **11**, 2079–2107 (2010)
5. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**(3), 1–27 (2011)
6. Chen, L.W., Rudnicky, A.: Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition. In: Proceedings of ICASSP, Rhodes Island, Greece (2023). https://doi.org/10.1109/ICASSP49357.2023.10095036
7. Egas-López, J.V., Svindt, V., Bóna, J., Hoffmann, I., Gosztolya, G.: Automated multiple sclerosis screening based on encoded speech representations. In: Proceedings of Interspeech, Dublin, Ireland, pp. 3003–3007 (2023)
8. Fan, Z., Li, M., Zhou, S., Xu, B.: Exploring wav2vec 2.0 on speaker verification and language identification. In: Proceedings of Interspeech, pp. 1509–1513 (2021)
9. Fara, S., Hickey, O., Georgescu, A., Goria, S., Molimpakis, E., Cummins, N.: Bayesian Networks for the robust and unbiased prediction of depression and its symptoms utilizing speech and multimodal data. In: Proceedings of Interspeech, Dublin, Ireland, pp. 1728–1732 (2023). https://doi.org/10.21437/Interspeech.2023-1709
10. Gosztolya, G., Tóth, L., Svindt, V., Bóna, J., Hoffmann, I.: Using acoustic deep neural network embeddings to detect multiple sclerosis from speech. In: Proceedings of ICASSP, Singapore, pp. 6927–6931 (2022)
11. Grosman, J.: Fine-tuned XLSR-53 large model for speech recognition in Hungarian (2021). https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-hungarian
12. Grósz, T., Porjazovski, D., Getman, Y., Kadiri, S., , Kurimo, M.: Wav2vec2-based paralinguistic systems to recognise vocalised emotions and stuttering. In: Proceedings of ACM Multimedia, Lisboa, Portugal, pp. 7026–7029 (2022)
13. Grósz, T., Virkkunen, A., Porjazovski, D., Kurimo, M.: Discovering relevant subspaces of BERT, Wav2Vec 2.0, ELECTRA and ViT embeddings for humor and mimicked emotion recognition with integrated gradients. In: Proceedings of MuSe, Ottawa, Canada, pp. 27–34 (2023). https://doi.org/10.1145/3606039.3613102
14. Hajduska-Dér, B., Kiss, G., Sztahó, D., Vicsi, K., Simon, L.: The applicability of the Beck Depression Inventory and Hamilton Depression Scale in the automatic recognition of depression based on speech signal processing. Front. Psychiat. **13**, 879896 (2022). https://doi.org/10.3389/fpsyt.2022.879896

15. Huckvale, M., Beke, A., Ikushima, M.: Prediction of sleepiness ratings from voice by man and machine. In: Proceedings of Interspeech, Shanghai, China, pp. 4571–4575 (2020)
16. Ivanova, O., Martínez-Nicolás, I., Meilán, J.J.G.: Speech changes in old age: methodological considerations for speech-based discrimination of healthy ageing and alzheimer's disease. Int. J. Lang. Commun. Disord. **59**(1), 13–37 (2023)
17. Jenei, A.Z., Kiss, G., Sztahó, D.: Detection of speech related disorders by pretrained embedding models extracted biomarkers. In: Proceedings of SPECOM, Gurugram, India, pp. 279–289 (2022)
18. Kiss, G., Tulics, M.G., Sztahó, D., Vicsi, K.: Language independent detection possibilities of depression by speech. In: Proceedings of NoLISP, pp. 103–114 (2016)
19. Klumpp, P., et al.: The phonetic footprint of Parkinson's disease. Comput. Speech Lang. **72**, 101321 (2022)
20. Kodali, M., Kadiri, S.R., Alku, P.: Classification of vocal intensity category from speech using the wav2vec2 and whisper embeddings. In: Proceedings of Interspeech, pp. 4134–4138 (2023). https://doi.org/10.21437/Interspeech.2023-2038
21. Kondratenko, V., Karpov, N., Sokolov, A., Savushkin, N., Kutuzov, O., Minkin, F.: Hybrid dataset for speech emotion recognition in Russian language. In: Proceedings of Interspeech, pp. 4548–4552 (2023). https://doi.org/10.21437/Interspeech.2023-311
22. Kumar, N., Nasir, M., Georgiou, P., Narayanan, S.S.: Robust multichannel gender classification from speech in movie audio. In: Proceedings of Interspeech, San Francisco, CA, USA, pp. 2233–2237 (2016)
23. Mann, H.B., Whitney, D.R.: On a test of whether one of two random variables is stochastically larger than the other. Ann. Math. Stat. **18**(1), 50–60 (1947)
24. Mihajlik, P., Balog, A., Gráczi, T.E., Kohári, A., Tarján, B., Mády, K.: BEA-Base: a benchmark for ASR of spontaneous Hungarian. In: Proceedings of LREC, pp. 1970–1977 (2022)
25. Mirheidari, B., O'Malley, R., Blackburn, D., Christensen, H.: Identifying people with mild cognitive impairment at risk of developing dementia using speech analysis. In: Proceedings of ASRU (2023). https://doi.org/10.1109/ASRU57964.2023.10389623
26. Pepino, L., Riera, P., Ferrer, L.: Emotion recognition from speech using wav2vec 2.0 embeddings. In: Proceedings of Interspeech, Brno, Czechia, pp. 3400–3404 (2021). https://doi.org/10.21437/Interspeech.2021-703
27. Pérez-Toro, P., et al.: Alzheimer's detection from English to Spanish using acoustic and linguistic embeddings. In: Proceedings of Interspeech, pp. 2483–2487 (2022)
28. Schneider, S., Baevski, A., Collobert, R., Auli, M.: wav2vec: unsupervised pretraining for speech recognition. In: Proceedings of Interspeech, pp. 3465–3469 (2019)
29. Szirmai, I.: Neurológia. Medicina, Budapest (2006)
30. Thienpondt, J., Speksnijder, C.M., Demuynck, K.: Behavioral analysis of pathological speaker embeddings of patients during oncological treatment of oral cancer. In: Proceedings of Interspeech, pp. 3018–3022 (2023). https://doi.org/10.21437/Interspeech.2023-1868
31. Tóth, L., et al.: Automatic detection of mild cognitive impairment from spontaneous speech using ASR. In: Proceedings of Interspeech, Dresden, Germany, pp. 2694–2698 (2015)

32. Vaessen, N., Van Leeuwen, D.A.: Fine-tuning wav2vec2 for speaker recognition. In: Proceedings of ICASSP, pp. 7967–7971 (2021)
33. Warule, P., Mishra, S.P., Deb, S.: Significance of voiced and unvoiced speech segments for the detection of common cold. Signal Image Video Process. **17**, 1785–1792 (2023)

Cross-Cultural Automatic Depression Detection Based on Audio Signals

Danila Mamontov[1,2(✉)], Sebastian Zepf[3], Alexey Karpov[4], and Wolfgang Minker[1]

[1] Ulm University, Ulm, Germany
{danila.mamontov,wolfgang.minker}@uni-ulm.de
[2] ITMO University, St. Petersburg, Russia
[3] Stuttgart, Germany
sebastian@zepf.info
[4] St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), St. Petersburg, Russia
karpov@iias.spb.su

Abstract. Depression is a frequently occurring mental health disorder globally, and early detection is critical for effective treatment. In this paper, we explore the effectiveness of machine learning techniques in cross-cultural depression detection using audio signals from Chinese and English-speaking populations. We investigate the influence of temporal context length, feature sets, and classifiers on classification performance across two single and two cross-corpus settings. Our results show that hand-crafted features offer advantages in single and combined dataset settings, while deep learning-based features, particularly from emotion recognition tasks, demonstrate superior cross-dataset generalization. The optimal length of a temporal context strongly depends on the specific dataset. These findings highlight the importance of considering dataset-specific characteristics and feature selection in developing reliable and culturally adaptable models for depression detection. In the cross-corpus settings on MENHIR and CMDC datasets, we obtained the best F1 scores of 0.77 and 0.63, respectively. Future research should focus on enhancing model performance and data accessibility to ensure effective inclusion across diverse populations, ultimately contributing to better mental health outcomes globally.

Keywords: Depression recognition · Deep embeddings · OpenSMILE · Cross-corpus analysis · CMDC · MENHIR

1 Introduction

Major Depressive Disorder (MDD) is a serious mental disease characterised by prolonged periods of low mood, loss of interest in pleasures, and reduced energy.

S. Zepf—Independent Researcher.

A. Karpov and V. Delić (Eds.): SPECOM 2024, LNAI 15299, pp. 309–323, 2025.
https://doi.org/10.1007/978-3-031-77961-9_23

This condition may negatively affect all aspects of a person's life, including their emotional, physical, and social well-being. Even the life expectancy of people with depression is 7.9 years shorter than that of others [21]. Meanwhile, the World Health Organization (WHO) reported that in 2023 about 5% of the adult population in the world suffered from depression [29].

The duration and effectiveness of treatment depend largely on the timely detection of MDD [11]. However, root-causes such as unawareness of those affected [33] or concealment because of discrimination fear [3] as well as insufficient availability and access to mental health services [26] are counterproductive for early detection and treatment.

To tackle the problem of early depression detection, various signals are considered in existing works. One of them is the speech signal. Acoustic speech characteristics of a person with a depressive disorder differ significantly from those of a healthy individual [2,7]. This finding suggests that it is possible to develop systems for automatic MDD detection through voice recordings, as the acoustic description of a speech signal gives us insight into how people speak.

Given that depression affects people regardless of their origin and place of residence, it seems timely to develop tools that can work equally effectively irrespective of language and cultural differences to increase awareness and support early detection. However, available public data is sparse, and models are often language-dependent. For a general improvement of society's mental health, language-independent tools for depression detection are needed.

Modern Machine Learning (ML) techniques enable more and more possibilities to develop reliable and effective systems for the automatic recognition of depressive disorders based on speech. As a rule, when using ML for MDD detection, it is necessary to choose an appropriate classification algorithm, feature set and the size of a temporal context.

To address this need, we have performed a systematic comparison of different temporal contexts, feature sets, and classifiers to detect MDD. We have also explored cross-corpus applications using Chinese and English-speaking data, striving to cope with data sparsity and language dependency.

2 Related Work

In this section, we describe the state-of-the-art in acoustic-based automatic depression detection with a focus on temporal context, audio features and cross-cultural application.

2.1 Temporal Context

The length of the temporal context during depression prediction from a speech signal is one aspect that has been shown to have a high impact on detection performance. For example, Aloshban et al. investigated the possibilities of recognizing depression from short passages of speech, less than 10 s in length, and showed on the DAIC-WOZ dataset that the effectiveness of such systems is

comparable to systems that use entire recordings [1]. Meanwhile, Dumpala et al. showed on the same dataset that as the temporal context increases, recognition accuracy also increases [6]. The discrepancy between these two works may be explained by a different approach to dividing the signal into parts. In the first paper, the division took place into semantic parts, and in the second paper, the division occurred simply by time in seconds. He et al. took windows of 20 s and achieved a Root Mean Square Error (RMSE) of 9.0 for both the AVEC2013 and AVEC2014 datasets [13], while Niu et al. obtained lowest RMSEs of 9.5 and 9.13 by using a temporal context of two seconds for these two datasets, respectively [19]. Overall, the amount of context length required seems to be dependent on each specific dataset used, which poses a challenge for cross-dataset applications.

2.2 Audio Features

Most commonly, the main efforts when developing ML systems for MDD detection are concentrated around the search for the most informative features. With the evolution of ML, the types of used features also change. Different hand-crafted features played a key role and showed their effectiveness in many studies aiming to solve the MDD recognition task [8,14,24]. They represent domain knowledge and human expertise meticulously engineered from raw data to be informative for a specific task and have the advantage of interpretability. Recently, along with the success and rapid growth of neural networks (NN) that was especially driven by architectures such as CNN and RNN [13,15,31], heuristic features have become increasingly used, namely coefficients from the last layers of NNs, called embeddings. Their main drawback is the impossibility of interpretation in contrast to hand-crafted ones. Recently, with the development of transformers, their success is only being consolidated, which may also be beneficial for cross-cultural applications [16,25,32].

2.3 Cross-Cultural Application

Depression detection from speech is a widely studied topic [12,18,30]. Existing work shows that speech analysis for depression detection is a promising approach, but overcoming language and cultural barriers is crucial for providing equitable mental healthcare [9,17]. The lack of training data for each certain language creates the need to explore the possibilities of cross-dataset applications. Further, the difficulties in collecting, storing and sharing mental health-related data, and other ethical issues limit the possibilities of increasing the data available for analysis. There are only a few existing datasets, especially when focusing on a specific language.

The DAIC-WOZ is the most popular dataset [10]. It was used for the Audio/Visual Emotion Challenges (AVEC) in 2016, 2017, and 2019 [22,23,27]. The AViD-Dataset was used for AVEC in 2013 [28]. Some other available datasets include CMDC, MODMA, and MENHIR [4,5,34]. The most common languages are English (DAIC-WOZ, MENHIR), Chinese(CMDC, MODMA) and German (AViD).

In this work, we aim to push language and culture-independent automatic detection of depression by building upon acoustic information from the speech signal. In particular, we investigate the application of trained classifiers on a foreign-language dataset. This approach aims to identify methods that can enhance the development of classifiers capable of reliable performance across diverse populations and to address the limited amount of available data.

3 Datasets Description

In this section, we describe two datasets that have been used in our work, namely the Chinese Multimodal Depression Dataset (CMDC) and the Mental Health Monitoring through Interactive Conversations (MENHIR). Datasets represent Chinese and English, respectively, and contain voice recordings from subjects with and without diagnosed MDD.

3.1 CMDC

The CMDC dataset has been created to address the challenge of undiagnosed depression in China [34]. It includes raw audio recordings from a semi-structured interview study with diagnosed individuals. CMDC comprises recordings from 78 subjects, including 26 with MDD and 52 without. It also contains extracted visual features from 45 subjects but without original videos included. The visual features are not considered in this work. All dialogues are in Chinese.

3.2 MENHIR

The MENHIR dataset has been collected within the MENHIR project which aims to help people with mental illness by means of conversational technologies [5]. It contains audio recordings in English from 51 subjects. Among them, 31 belonged to the group with active MDD, while the remaining 20 were part of a control group with no history of MDD issues. The interview consisted of 14 questions taken from the Warwick-Edinburgh Mental Wellbeing Scale.

4 Methodology

In this section, we describe our approach to examine the impact of the temporal context lengths, feature sets, and classifiers on the classification performance in the same and cross-culture settings. The schematic representation of our approach is shown in Fig. 1. We extracted all features by applying sliding windows with different lengths. Then we trained classifiers on each feature set separately. Finally, we tested all models using stratified 10-fold cross-validation in four evaluation settings.

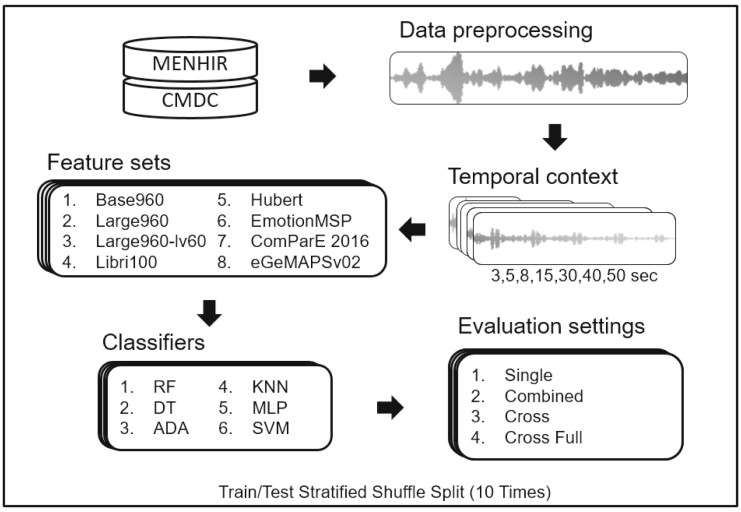


Fig. 1. The pipeline of the proposed approach.

4.1 Data Preprocessing

The CMDC dataset includes preprocessed audio files. The speaker diarisation has already been done by the authors of the data. All dialogues in the dataset are divided into separate files based on the twelve questions asked to the participants in the experiment. Therefore, we did not make any changes to the original audio files before extracting the features.

Regarding the MENHIR dataset, we used the markup provided by the authors to perform the diarisation. After that, we extracted all the features in the same way as with the CMDC dataset.

4.2 Temporal Context

We used sliding windows with different window lengths for feature extraction. Taking into account the distribution of record lengths from both datasets, we took temporal context length of **3, 5, 8, 15, 30, 40, and 50 s**, which allows us to keep all subjects represented by at least one audio record within our experiment. Window shifts for all window lengths were equal to half the window lengths.

4.3 Classifiers

We took six well-established classification algorithms: Random Forest (**RF**), Decision Tree (**DT**), AdaBoost (**ADA**), k-Nearest Neighbors (**KNN**), Support Vector Machine (**SVM**), Multi-Layer Perceptron (**MLP**). For all the algorithms, implementations from the scikit-learn library were used [20]. While we primarily utilized default hyperparameter settings, some specific choices were made as follows:

– RF with max_depth=5 and n_estimators=10
– DT with max_depth=5
– ADA with n_estimators=50
– KNN with n_neighbors=5
– SVM with kernel="rbf"
– MLP with max_iter=200, alpha=0.5

4.4 Feature Sets

We chose the feature sets listed in Table 1 for our study. We included deep learning embeddings from models taken from HuggingFace[1], along with hand-crafted features extracted through the openSMILE toolkit[2]. The deep learning models included Base960, Large960, Large960-lv60, Libri100, and LargeHuBert, which were trained on the Librispeech dataset for Automatic Speech Recognition (ASR). They represent the speech signal using learned features suitable for various ASR tasks in different languages including English and Chinese. The EmotionMSP model is based on the Wav2Vec2-Large-Robust model and is trained for emotion recognition. ComParE 2016 and eGeMAPSv02 are hand-crafted feature sets specifically designed for audio analysis. We applied PCA on ComParE 2016 for a dimensionality reduction, saving 95% of the initial variance.

Table 1. Feature sets used in our study with their full name, short name, and size.

Full Name	Short Name	Size
facebook/wav2vec2-base-960h	Base960	768
facebook/wav2vec2-large-960h	Large960	1024
facebook/wav2vec2-large-960h-lv60-self	Large960-lv60	1024
patrickvonplaten/wavlm-libri-clean-100h-base-plus	Libri100	768
facebook/hubert-large-ls960-ft	LargeHuBert	768
audeering/wav2vec2-large-robust-12-ft-emotion-msp-dim	EmotionMSP	1024
openSMILE ComParE_2016 Functionals	ComParE 2016	6373*
openSMILE eGeMAPSv02 Functionals	eGeMAPSv02	88

*PCA is applied, retaining 95% of the variance.

4.5 Evaluation

Combining seven context lengths, eight feature sets, and six classifiers results in 336 combinations. We tested all of them using stratified 10-fold cross-validation in two single and two cross-corpus settings. Since we aim to identify promising

[1] https://huggingface.co/models Accessed: 2024-05-24.
[2] https://scikit-learn.org/stable Accessed: 2024-05-24.

approaches for cross-cultural depressive disorder detection, we focus on the models' performance in identifying the depressive disorder class, while not taking into account the effectiveness of recognizing people without a depressive disorder in this work. We took the F1 score as a performance metric, which balances precision and recall. Final evaluations are done on the subject level, by averaging the predictions from all windows of a certain subject. If the average value exceeds 0.5, then the positive MDD class is assigned.

To analyze models in both the same-culture and cross-cultural cases, using data from the English and Chinese datasets, we have defined the following settings:

- **Single** considers test and training samples coming from the same dataset.
- **Combined** mixes training parts of both datasets, followed by separate testing on each testing part of the individual dataset.
- **Cross** involves training the model on the training part of one dataset and testing it on the testing part of another.
- **Cross Full** utilizes the training portion of one dataset and the other entire dataset for testing.

Hence, it is feasible to directly compare the outcomes of Single, Combined, and Cross with each other. The Full Cross setting provides additional insights into model performance with a larger test set, reflecting real-world and cross-cultural settings with potentially increased data availability.

4.6 Baseline

As benchmarks, we consider a dummy classifier that always predicts the true class and results reported on CMDC and MENHIR from other researchers, as presented in Table 2. There are no studies for MENHIR that used acoustic features for depression prediction. Zubiaga and Justo leveraged linguistic features and obtained the F1 score of 0.93 [35]. The highest F1 score of 0.94 reported by the authors of the CMDC dataset was achieved using a combination of acoustic and linguistic features [34].

5 Results

In this section, we present the results of our experiments. We start by considering all three parameters separately, averaging the F1 scores across the two remaining parameters, and at the end, we present the best combinations. In Fig. 2-4, the bottom black horizontal line of the box plot represents the minimum value, the first, second (median), and third lines within the box correspond to the 25th, 50th, and 75th percentiles, respectively, the top line represents the maximum value, and the mean F1 score is indicated by dots. When discussing the values of the F1 score in the further course of this work, we refer to the mean F1 score.

Table 2. Benchmarks on CMDC and MENHIR datasets in four test settings.

Setting	Dataset	Classifier	F1
Single	CMDC	Dummy	0.50
		Baseline	0.94
	MENHIR	Dummy	0.57
		Baseline	0.93
Combined	CMDC	Dummy	0.50
	MENHIR	Dummy	0.57
Cross	CMDC	Dummy	0.50
	MENHIR	Dummy	0.57
Cross Full	CMDC	Dummy	0.48
	MENHIR	Dummy	0.56

5.1 Temporal Context

Figure 2 presents the performance of different context lengths. We can observe that on the CMDC dataset, the best context lengths are 30 and 40 s. They achieved the best results in all four settings namely 0.67, 0.70, 0.47, and 0.50 F1 scores in Single, Combined, Cross, and Cross Full settings respectively.

For MENHIR the best context length in Single and Combined settings is 5 s with 0.84 and 0.81 F1 scores respectively. In Cross and Cross Full settings the longer context length of 40 s worked out better achieving 0.38 and 0.40 F1 scores respectively.

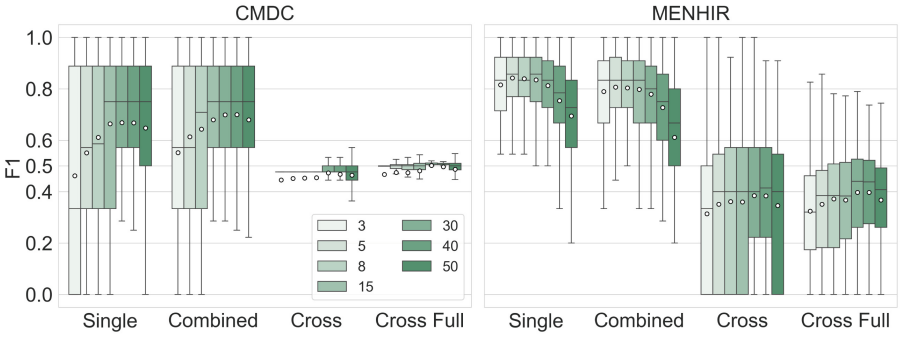


Fig. 2. Context lengths' performance, averaged by feature set and classifier, on CMDC and MENHIR datasets using the F1 score.

These findings indicate that for the CMDC dataset, a window length of approximately 30–40 s is the most effective choice. In contrast, for the MENHIR dataset, the optimal context length varies significantly depending on the specific

usage setting. On the contrary, the results in the Cross and Cross Full settings are worse than those of the dummy classifier. This suggests that when an unsuitable combination of feature set and classifier is used, the context length does not generally offer the opportunity to improve performance for cross-cultural application.

5.2 Features

Figure 3 illustrates the performance of various feature sets. Several notable patterns emerge from the data. In the Single setting, hand-crafted features consistently achieved the highest performance across both datasets. Specifically, the eGeMAPSv02 feature set attained an F1 score of 0.95 on the CMDC dataset, while the ComParE 2016 feature set reached the highest F1 score of 0.89 on the MENHIR dataset. In the Combined setting, eGeMAPSv02 outperformed all other feature sets, with F1 scores of 0.95 on CMDC and 0.84 on MENHIR.

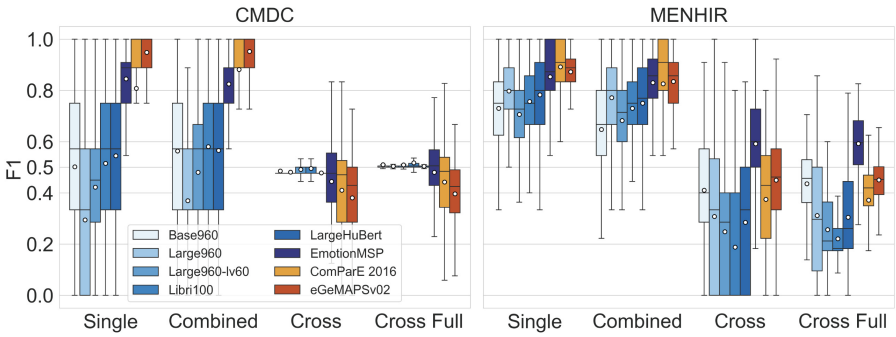


Fig. 3. Feature sets' performance, averaged by window length and classifier, on CMDC and MENHIR datasets using the F1 score.

Conversely, results from the Cross and Cross Full tests indicated an opposite trend, where hand-crafted features performed worse. Instead, the EmotionMSP feature set achieved the best F1 score of 0.59 on the MENHIR dataset. Additionally, the Libri100 feature set attained the highest F1 scores of 0.50 in the Cross setting and 0.52 in the Cross Full setting on CMDC.

This contrast highlights the varying effectiveness of feature sets depending on the setting, suggesting that the choice of feature set is crucial for optimal performance across different contexts, but we can note the potential of EmotionMSP since it has significantly higher results in Cross and Cross Full settings on the MENHIR dataset.

5.3 Classifiers

Figure 4 presents the performance of different classifiers. We can conclude that ADA and MLP showed the most prospective performance. ADA achieved the

highest scores on both datasets in Single and Combined settings. MLP obtained the best F1 scores of 0.43 and 0.45 in Cross and Cross Full settings on MENHIR. Meanwhile, on the CMDC dataset RF showed the best F1 score of 0.48 in the Cross setting and KNN with the 0.51 F1 score outperformed other classifiers in the Cross Full setting.

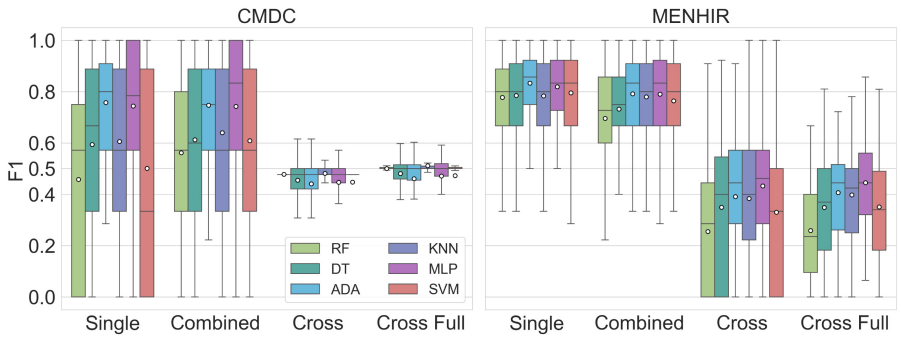


Fig. 4. Classifiers' performance, averaged by window length and feature set, on CMDC and MENHIR datasets using the F1 score.

5.4 Best Configurations

Table 3 presents the optimal combinations of window lengths, feature sets, and classifiers across all four settings, as determined by their F1 scores. Several noteworthy trends are observed.

Window length: In both the Single and Combined settings, the most common window lengths among the top-performing combinations are 3, 5, 8, 15 s. In the Cross and Cross Full settings, the optimal context length varies depending on the dataset. For the CMDC dataset, the best window lengths are 30 and 40 s, whereas for the MENHIR dataset, a window length of 3 s suffices.

Features: In the Combined setting, only hand-crafted feature sets appear among the top-performing combinations. Conversely, in the Single setting, hand-crafted and EmotionMSP feature sets are represented among the best-performing combinations. We also note that EmotionMSP appeared to be the best feature set in Cross and Cross Full settings for both datasets.

Classifiers: In both the Single and Combined settings, numerous combinations yielded the highest F1 scores, precluding the selection of a single best configuration. When examining the classifiers, it is evident that only the MLP, SVM, and ADA classifiers are part of the best-performing combinations in the Combined setting. In contrast, the Single setting includes all six classifiers among its top-performing combinations. SVM appears to be the best among all classifiers in Cross and Cross Full settings.

Table 3. Best found combinations of context length, feature set, and classifier by F1 metric.

Setting	Test Dataset	Window length (sec)	Feature set	Classifier	F1
Single	CMDC	15	ComParE 2016	MLP	1.00
		30	ComParE 2016	MLP	1.00
	MENHIR	3	LargeHuBert	SVM	0.94
		3	ComParE 2016	DT	0.94
		3	ComParE 2016	KNN	0.94
		5	EmotionMSP	RF	0.94
		5	EmotionMSP	SVM	0.94
		5	ComParE 2016	DT	0.94
		5	ComParE 2016	ADA	0.94
		5	ComParE 2016	KNN	0.94
		8	EmotionMSP	RF	0.94
		8	EmotionMSP	ADA	0.94
		8	EmotionMSP	MLP	0.94
		8	EmotionMSP	SVM	0.94
		8	ComParE 2016	KNN	0.94
		15	EmotionMSP	ADA	0.94
		15	ComParE 2016	KNN	0.94
Combined	CMDC	3	ComParE 2016	MLP	0.99
		3	ComParE 2016	SVM	0.99
		3	eGeMAPSv02	ADA	0.99
		5	ComParE 2016	MLP	0.99
		5	ComParE 2016	SVM	0.99
		5	eGeMAPSv02	ADA	0.99
		8	ComParE 2016	MLP	0.99
		8	ComParE 2016	SVM	0.99
		8	eGeMAPSv02	ADA	0.99
		15	ComParE 2016	MLP	0.99
		15	ComParE 2016	SVM	0.99
		15	eGeMAPSv02	ADA	0.99
		30	ComParE 2016	MLP	0.99
	MENHIR	3	eGeMAPSv02	SVM	0.97
Cross	CMDC	30	EmotionMSP	SVM	0.60
	MENHIR	3	EmotionMSP	MLP	0.71
Cross Full	CMDC	40	EmotionMSP	SVM	0.63
	MENHIR	3	EmotionMSP	SVM	0.77

When moving from the Single to Combined setting the best performance increases for the MENHIR dataset, but slightly decreases for the CMDC dataset. It shows that extending the training set of CMDC with samples from MENHIR badly influences the models. This may be due to differences in language as well as the quality of the MENHIR dataset. On the other hand, expanding the training set of MENHIR with samples from CMDC made it possible to improve the quality of the models.

All the best combinations in each setting give better results than the corresponding benchmarks listed in Table 2.

6 Discussion

The results of our work show that existing models are not suitable for recognizing depression in different contexts and languages without preliminary fine-tuning of models. We found that models trained on the CMDC dataset have the potential to predict MDD on MENHIR without preliminary fine-tuning. However, models trained on the MENHIR dataset did not work well on average, except those using the EmotionMSP feature set, and only some of the best combinations achieved promising results. This may be due to various factors, including the limitations of the datasets used for training. One notable challenge in this regard was related to the MENHIR dataset. The target and control groups in this dataset had different acoustic characteristics, which were not caused by differences in the subjects' mental state but by differences in the quality of the audio recording equipment. This may partly explain the low performance in the Cross and Cross Full settings on the CMDC dataset.

An important consideration when evaluating the datasets used in this study is their specificity to clinical interviews centred around the condition being investigated. Unlike datasets that cover a broader spectrum of general speech applications, these datasets are derived from interactions that are inherently focused on the clinical context. This specificity means that both the semantics of the responses and the acoustic features encoded in the speech signals are likely to be influenced by the subject matter of the interviews. In general, this is a domain's problem, since data collection from people with mental illnesses is fraught with ethical limitations and it cannot be carried out completely in out-of-laboratory conditions.

7 Conclusion

In conclusion, our study shows the importance of considering dataset-specific characteristics when designing ML models for cross-cultural depression detection. While longer temporal contexts improve the classification quality on CMDC and the best results on MENHIR are achieved with shorter context lengths. Hand-crafted features offer advantages in Single and Combined settings, and EmotionMSP shows superior for generalization in Cross-corpus settings. ADA and MLP classifiers have proven effectiveness in Single and Combined settings,

but Cross-corpus applications may require more nuanced approaches. Therefore, further research efforts should focus on improving model performance considering further datasets and increasing data accessibility to ensure inclusion across diverse populations.

Acknowledgments. This research was partially supported by RSF in the framework of the project No. 22-11-00321 (A. Karpov).

References

1. Aloshban, N., Esposito, A., Vinciarelli, A.: Detecting depression in less than 10 seconds: impact of speaking time on depression detection sensitivity. In: Proceedings of the 2020 International Conference on Multimodal Interaction, ICMI '20, pp. 79–87. Association for Computing Machinery, New York (2020). https://doi.org/10.1145/3382507.3418875
2. Alpert, M., Pouget, E.R., Silva, R.R.: Reflections of depression in acoustic measures of the patient's speech. J. Affect. Disord. **66**(1), 59–69 (2001). https://doi.org/10.1016/S0165-0327(00)00335-9. https://www.sciencedirect.com/science/article/pii/S0165032700003359
3. Brohan, E., Gauci, D., Sartorius, N., Thornicroft, G.: Self-stigma, empowerment and perceived discrimination among people with bipolar disorder or depression in 13 European countries: the GAMIAN–Europe study. J. Affect. Disord. **129**(1), 56–63 (2011). https://doi.org/10.1016/j.jad.2010.09.001. https://www.sciencedirect.com/science/article/pii/S0165032710005690
4. Cai, H., et al.: A multi-modal open dataset for mental-disorder analysis. Sci. Data **9**(1), 178 (2022). https://doi.org/10.1038/s41597-022-01211-x. https://www.nature.com/articles/s41597-022-01211-x
5. Callejas Carrión, Z., Benghazi, K., Noguera, M., Torres Barañano, M.I., Justo Blanco, R.: MENHIR: mental health monitoring through interactive conversations (2019). https://doi.org/10.26342/2019-63-15. http://rua.ua.es/dspace/handle/10045/96617
6. Dumpala, S.H., Rodriguez, S., Rempel, S., Sajjadian, M., Uher, R., Oore, S.: Detecting depression with a temporal context of speaker embeddings (2022)
7. France, D., Shiavi, R., Silverman, S., Silverman, M., Wilkes, M.: Acoustical properties of speech as indicators of depression and suicidal risk. IEEE Trans. Biomed. Eng. **47**(7), 829–837 (2000). https://doi.org/10.1109/10.846676. https://ieeexplore.ieee.org/abstract/document/846676
8. Gong, Y., Poellabauer, C.: Topic modeling based multi-modal depression detection. In: Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, AVEC 2017, pp. 69–76. Association for Computing Machinery, New York (2017). https://doi.org/10.1145/3133944.3133945
9. Gotlib, I.H., Hammen, C.L.: Handbook of Depression, 2nd edn. Guilford Press (2008)
10. Gratch, J., et al.: The distress analysis interview corpus of human and computer interviews. In: Calzolari, N., et al. (eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), pp. 3123–3128. European Language Resources Association (ELRA), Reykjavik (2014). http://www.lrec-conf.org/proceedings/lrec2014/pdf/508_Paper.pdf

11. Halfin, A.: Depression: the benefits of early and appropriate treatment. Am. J. Manag. Care **13**(4 Suppl), S92-97 (2007)
12. Han, M.M., et al.: Automatic recognition of depression based on audio and video: a review. World J. Psychiatry **14**(2), 225–233 (2024). https://doi.org/10.5498/wjp.v14.i2.225. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10921287/
13. He, L., Cao, C.: Automated depression analysis using convolutional neural networks from speech. J. Biomed. Inf. **83**, 103–111 (2018). https://doi.org/10.1016/j.jbi.2018.05.007. https://www.sciencedirect.com/science/article/pii/S153204641830090X
14. He, L., Jiang, D., Sahli, H.: Multimodal depression recognition with dynamic visual and audio cues. In: 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 260–266 (2015). https://doi.org/10.1109/ACII.2015.7344581. https://ieeexplore.ieee.org/abstract/document/7344581, iSSN: 2156-8111
15. Kaya, H., et al.: Predicting depression and emotions in the cross-roads of cultures, para-linguistics, and non-linguistics. In: Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop, AVEC 2019, pp. 27–35. Association for Computing Machinery, New York (2019). https://doi.org/10.1145/3347320.3357691
16. Lam, G., Dongyan, H., Lin, W.: Context-aware deep learning for multimodal depression detection. In: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3946–3950 (2019). https://doi.org/10.1109/ICASSP.2019.8683027. https://ieeexplore.ieee.org/abstract/document/8683027, iSSN: 2379-190X
17. Lehti, A., Hammarström, A., Mattsson, B.: Recognition of depression in people of different cultures: a qualitative study. BMC Family Pract. **10**(1), 53 (2009). https://doi.org/10.1186/1471-2296-10-53
18. Meng, H., Huang, D., Wang, H., Yang, H., AI-Shuraifi, M., Wang, Y.: Depression recognition based on dynamic facial and vocal expression features using partial least square regression. In: Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge, AVEC 2013, pp. 21–30. Association for Computing Machinery, New York (2013). https://doi.org/10.1145/2512530.2512532
19. Niu, M., Tao, J., Liu, B., Huang, J., Lian, Z.: Multimodal spatiotemporal representation for automatic depression level detection. IEEE Trans. Affect. Comput. **14**(1), 294–307 (2023). https://doi.org/10.1109/TAFFC.2020.3031345. https://ieeexplore.ieee.org/abstract/document/9226102
20. Pedregosa, F., et al.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
21. Pratt, L.A., Druss, B.G., Manderscheid, R.W., Walker, E.R.: Excess mortality due to Depression and Anxiety in the United States: results from a nationally representative survey. General Hosp. Psychiat. **39**, 39–45 (2016). https://doi.org/10.1016/j.genhosppsych.2015.12.003. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5113020/
22. Ringeval, F., Schuller, B., Valstar, M., Cummins, N., Cowie, R., Pantic, M.: AVEC'19: audio/visual emotion challenge and Workshop, pp. 2718–2719 (2019). https://doi.org/10.1145/3343031.3350550
23. Ringeval, F., ET AL.: AVEC 2017: real-life depression, and affect recognition workshop and challenge. In: Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, AVEC '17, pp. 3–9. Association for Computing Machinery, New York (2017). https://doi.org/10.1145/3133944.3133953

24. Sidorov, M., Minker, W.: Emotion recognition and depression diagnosis by acoustic and visual features: a multimodal approach. In: Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, AVEC '14, pp. 81–86. Association for Computing Machinery, New York (2014). https://doi.org/10.1145/2661806.2661816
25. Sun, H., et al.: Multi-modal adaptive fusion transformer network for the estimation of depression level. Sensors **21**(14), 4764 (2021). https://doi.org/10.3390/s21144764. https://www.mdpi.com/1424-8220/21/14/4764
26. Thomas, K.C., Ellis, A.R., Konrad, T.R., Holzer, C.E., Morrissey, J.P.: County-level estimates of mental health professional shortage in the United States. Psychiat. Serv. **60**(10), 1323–1328 (2009). https://doi.org/10.1176/ps.2009.60.10.1323. https://ps.psychiatryonline.org/doi/full/10.1176/ps.2009.60.10.1323
27. Valstar, M., et al.: AVEC 2016: depression, mood, and emotion recognition workshop and challenge. In: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, AVEC '16, pp. 3–10. Association for Computing Machinery, New York (2016). https://doi.org/10.1145/2988257.2988258
28. Valstar, M., et al.: AVEC 2013: the continuous audio/visual emotion and depression recognition challenge. In: Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge, AVEC 2013, pp. 3–10. Association for Computing Machinery, New York (2013). https://doi.org/10.1145/2512530.2512533
29. WHO: Depressive disorder WHO (depression) (2023). https://www.who.int/news-room/fact-sheets/detail/depression
30. Wu, P., Wang, R., Lin, H., Zhang, F., Tu, J., Sun, M.: Automatic depression recognition by intelligent speech signal processing: a systematic survey. CAAI Trans. Intell. Technol. **8**(3), 701–711 (2023). https://doi.org/10.1049/cit2.12113. https://onlinelibrary.wiley.com/doi/abs/10.1049/cit2.12113
31. Yang, L., Jiang, D., Han, W., Sahli, H.: DCNN and DNN based multimodal depression recognition. In: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 484–489 (2017). https://doi.org/10.1109/ACII.2017.8273643. https://ieeexplore.ieee.org/abstract/document/8273643. iSSN: 2156-8111
32. Yin, F., Du, J., Xu, X., Zhao, L.: Depression detection in speech using transformer and parallel convolutional neural networks. Electronics **12**(2), 328 (2023). https://doi.org/10.3390/electronics12020328. https://www.mdpi.com/2079-9292/12/2/328
33. Yu, Y., et al.: Recognition of depression, anxiety, and alcohol abuse in a Chinese rural sample: a cross-sectional study. BMC Psychiat. **16**(1), 93 (2016). https://doi.org/10.1186/s12888-016-0802-0
34. Zou, Bet al.: Semi-structural interview-based Chinese multimodal depression corpus towards automatic preliminary screening of depressive disorders. IEEE Trans. Affect. Comput. 1–16 (2022). https://doi.org/10.1109/TAFFC.2022.3181210. https://ieeexplore.ieee.org/document/9793717/algorithms
35. Zubiaga, I., Justo, R.: Multimodal feature evaluation and fusion for emotional well-being monitorization. In: Pinho, A.J., Georgieva, P., Teixeira, L.F., Sánchez, J.A. (eds.) Pattern Recognition and Image Analysis, pp. 242–254. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-04881-4_20

Depression Classification Using Token Merging-Based Speech Spectrotemporal Transformer

Lokesh Kumar(✉), Kumar Kaustubh, and S. R. Mahadeva Prasanna

Indian Institute of Technology Dharwad, Dharwad, India
{EE23MT012,221022003,prasanna}@iitdh.ac.in

Abstract. This paper introduces a novel approach for depression classification, utilizing multimodal token merging (ToMe) within a speech spectrotemporal transformer framework. The model's efficacy is evaluated with log-mel spectrograms and autocorrelation tempograms extracted from depressed and non depressed speech. The results demonstrate the effectiveness of ToMe when integrated with attention mechanisms of audio spectrogram transformer (AST) models, such as AST and data efficient image transformer (DeiT) encoders. This underscores the importance of the token pruning mechanism utilized in the study. Additionally, a multimodal dual-channel architecture is introduced, featuring two distinct feature modalities extracted from speech: spectrograms and autocorrelation tempograms. The novel ToMe dual-channel AST and ToMe dual-channel AST with DeiT encoder models demonstrate remarkable performance on two different datasets, namely the EATD-Corpus (Chinese) and DAIC-WoZ (English), providing promising results for depression detection.

Keywords: Log-mel spectrogram · Autocorrelation tempogram · Vision transformer · Token merging (ToMe)

1 Introduction

Depression is a psychological disorder requiring swift intervention and assessment. It may manifest as prolonged episodes of sadness, emptiness, or irritability, alongside noticeable mental and physical changes that persist for at least two weeks, severely affecting a person's ability to carry out daily activities [3]. With the escalating interest in affective computing applications, the task of depression detection [8] from speech [5] continues to be a challenging yet crucial endeavor. Speech-based analysis [1] has emerged as a valuable modality for understanding emotional states and mental health conditions. The human voice carries nuanced information, including pitch, tone, and rhythm, which can reflect the speaker's emotional well-being. Additionally, it offers a cost-efficient and remote method of evaluation, while ensuring the patient's identity and privacy are safeguarded, making it an excellent option for this purpose. Previous research has

A. Karpov and V. Delić (Eds.): SPECOM 2024, LNAI 15299, pp. 324–335, 2024.
https://doi.org/10.1007/978-3-031-77961-9_24

demonstrated the efficacy of utilizing speech features for emotion recognition [9], paving the way for exploring their application in depression detection. Currently, diagnosis heavily depends on methods like interviews and surveys, which are susceptible to biases and potential misdiagnoses [11]. Speech analysis offers a promising alternative as it directly reflects emotional and cognitive states, providing a more objective approach [17].

To extract meaningful features from the speech data, log-mel spectrograms have been employed in this work. It is commonly used in speech processing. However, such traditional spectral representations may not fully capture the temporal dynamics and rhythmic patterns present in depressed speech. To address this limitation, autocorrelation tempograms [14] have been utilized as an additional feature representation. Tempograms, commonly used in music analysis [14] for rhythm and tempo estimation, offer insights into the temporal structure of audio signals. By computing the autocorrelation of speech frames over time, a temporal representation that highlights recurring patterns and rhythmic elements in speech is obtained. To process and understand the complex patterns within the extracted speech features, we turn to state-of-the-art vision transformers architectures [7] that have shown exceptional performance in vision related tasks. Their adaptation to speech-based analysis for depression classification has been put forward in this work. The proposed methodology employs a dual-channel audio spectrogram transformer (AST) [12] architecture to concurrently process log-mel spectrograms and autocorrelation tempograms, thereby capturing both spectral and temporal dependencies in speech associated with depression. The integration of token merging in attention mechanisms across successive layers facilitates a more robust feature representation and also makes the model computationally faster, as the number of tokens keeps reducing, promoting enhanced feature extraction. For this purpose, a dual-channel approach with Token Merging (ToMe) [4] is proposed, where the ToMe is applied in AST and AST with data-efficient image transformer (DeiT) [30] encoders. This approach aims to combine the two features for better detection of depression-indicative patterns.

1.1 Related Works

Researchers have utilized different methods to effectively categorize speech samples as either indicating depression or not. Various spectral, glottal, and prosodic features have proven to be effective in this task [17]. Machine learning (ML) classifiers such as decision trees, BayesNet, and random forest have been used with features like MFCCs, jitter, and cepstral coefficients to perform the classification [32]. Low-Level Descriptors (LLDs) and their statistical measures have been employed as features in ML algorithms such as support vector machines (SVM), logistic regression (LR), and gaussian mixture models (GMM) to develop classification systems [34]. Features based on rhythm, such as rhythm formants, which are commonly used to study variations in speaking style within a language, have also proven useful in the classification of depressed speech [18]. Apart from the ML-based classifiers, researchers have also come up with deep learning (DL) techniques such as convolutional neural networks (CNNs) and recurrent

neural networks (RNNs) for evaluating depression severity [35]. Furthermore, researchers have investigated deep learning architectures that utilize combinations of CNNs and RNNs applied to spectrograms [28] and a combination of transformers and CNNs [20]. Convolutional autoencoders [27] and deep learnt features from waveform or spectrograms also exists in literature [24]. In addition to ML and DL-based systems, end-to-end solutions such as DepAudioNet have also been created and documented in the field [21].

This paper makes two major contributions: firstly, it introduces a novel DL-based architecture that integrates the concepts of ToMe, AST, and DeiT encoders for classifying depressed speech. Secondly, it suggests using autocorelation tempograms alongside log-mel spectrograms to improve classification performance by capturing both temporal and rhythmic features: autocorelation tempograms for rhythm and log-mel spectrograms for spectral content.

The paper is structured as follows: Sect. 2 describes the proposed methodology and classification system. Section 3 describes about the datasets used, experimental settings, model architecture and outlines the obtained results. Finally, the work is concluded in Sect. 4, giving some future work directions.

2 Proposed Method

In this section, feature extraction steps along with the description of the proposed deep learning model and the end-to-end classification system have been discussed. The model parameters and exact experimental setting are presented in the next section.

2.1 Features Extraction

Log-mel spectrogram extraction involves converting speech signals into a visual representation that emphasizes frequency and time information. It is extracted from speech by collecting the frequency components in the Mel scale, followed by a logarithmic operation, and plotting it over time. Apart from that, the autocorrelation tempogram, typically used for music analysis, is employed as a feature. The process for extracting it for speech is as follows:

The raw speech signal is represented as a discrete-time waveform $x[n]$, where n represents the sample index. Given the magnitude spectrogram $X(f,t)$ of the speech signal, where f represents frequency and t represents time, the spectral flux $F(t)$ at time instant t is calculated as:

$$F(t) = \sum_{f} |X(f,t) - X(f,t-1)|^2 \tag{1}$$

Next, the onset strength envelope $O[n]$ is derived from the spectral flux [14] by smoothing it using a Gaussian filter. The onset strength envelope $O[n]$ is then divided into frames of fixed duration using a Hamming window. Let $O_k[n]$

represent the k-th frame of the onset strength envelope. Mathematically, this can be expressed as:

$$O_k[n] = O[n] \cdot w[n - k \cdot \Delta] \tag{2}$$

where $w[n]$ is the Hamming window function and Δ is the frame shift.

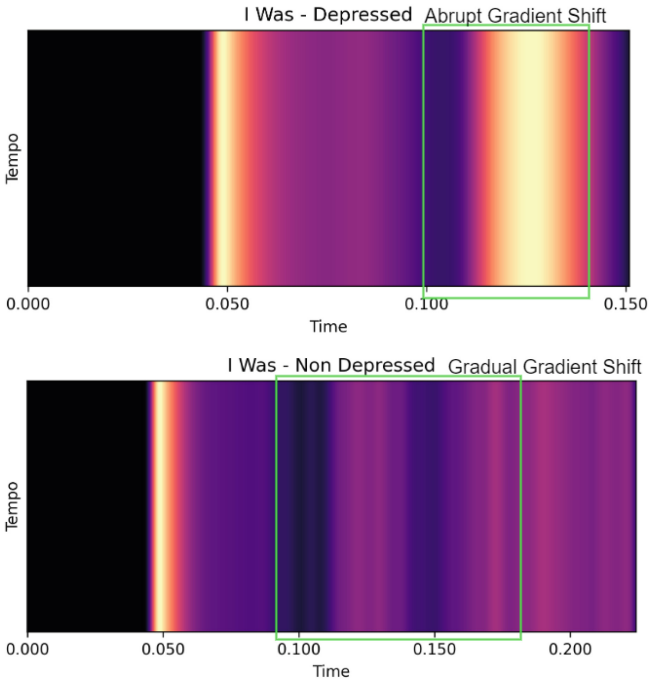


Fig. 1. Autocorrelation tempogram of phrase "I was" uttered by a depressed (top) and a non-depressed individual (bottom).

Finally, autocorrelation is computed and normalized for each frame $O_k[n]$ to measure the self-similarity or periodicity within the frame. The autocorrelation function $R_k[m]$ [14] for the k-th frame is calculated as:

$$R_k[m] = \sum_n O_k[n] \cdot O_k[n - m] \tag{3}$$

where m represents the lag or time shift. Figure 1 shows the autocorelation tempograms for a depressed and non-depressed individual uttering a common phrase, taken from the DAIC-WOZ dataset and the difference observed is marked inside the green box. Such abrupt gradient changes are observed across multiple sample pairs; however, this difference may not necessarily be associated with depression. It can be due to the different speech properties associated with various speakers, speaking rates, phonemes preceding and following the common utterance, etc. These features are extracted for all the speech samples and they serve as the input to the deep learning model, discussed in the next sub-section.

2.2 Dual-Channel Multimodal Fusion with Token Merging

To advance the feature extraction capabilities, a dual-channel multimodal classifier is proposed along with fusion mechanism involving ToMe across multiple layers. The model architecture, shown in Fig. 2, consists of two parallel processing streams, M_{spec} and M_{tempo}, corresponding to the spectrogram and tempogram channels, respectively. For the spectrogram channel, the input features have dimensions of size (768, 3, 16, 16), indicating a 3-channel input with a spatial resolution of 16×16 and processed through 12 encoder layers. The output layer of the spectrogram channel has dimensions of size (2, 768), signifying a linear transformation to a 2-class output. Similarly, for the tempogram channel, the input features also have dimensions of size (768, 3, 16, 16), with the same spatial resolution and the same number of encoder layers, i.e. 12. The output layer of the tempogram model is also identical to that of the spectrogram model, with dimensions size (2, 768).

In the dual-channel model, the combined features from both channels are concatenated and passed through an output classifier layer. The output classifier layer has dimensions size (2, 4), indicating a linear transformation from the concatenated feature space to a 2-class output and the AST produces feature representations for both channels: $F_{\text{spec},i}$ is the feature representation extracted by the AST for the spectrogram channel. $F_{\text{tempo},i}$ is the feature representation extracted by the AST for the tempogram channel. ToMe is introduced in each block and modified attention mechanism introduced by Meta Research [4] is used.

To integrate information across tokens within each layer, a token merging operation is applied with a reduction parameter r controlling the number of tokens per layer, the choice of optimal r value is chosen based on accuracy calculated on combined spectrogram and tempogram dataset split into training and validation set as shown in Fig. 3, enables efficient fusion of token-level information. $F'_{\text{spec},i}$ represents the merged token-level features for the spectrogram channel at layer i and is given as:

$$F'_{\text{spec},i} = M(F_{\text{spec},i}, r) \tag{4}$$

Similarly, $F'_{\text{tempo},i}$ represents the merged token-level features for the tempogram channel at layer i and is given as follows:

$$F'_{\text{tempo},i} = M(F_{\text{tempo},i}, r) \tag{5}$$

The output classifier combines the predictions from both the branches. It takes a concatenated representation of the outputs from the spectrogram and tempogram classifiers as input, resulting in a 4-dimensional feature vector, given by final joint feature representation, F_{joint}, constructed by concatenating or adding token-level features from each channel across all layers.

$$F_{\text{joint}} = \bigoplus_{i=1}^{N} F'_{\text{spec},i} \oplus \bigoplus_{i=1}^{N} F'_{\text{tempo},i} \tag{6}$$

This vector is then fed into a fully connected layer, which produces the final classification decision by mapping the combined features to a binary output space of size 2, enabling the model to intricately capture hierarchical and cross-layer relationships in both channels.

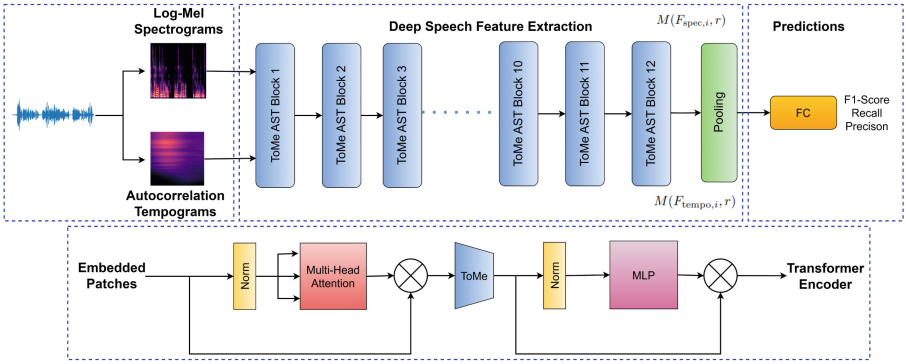


Fig. 2. Proposed Method. Raw speech signals serve as the initial input to the augmentation process. The variability in window sizes is used during log-mel spectrogram and autocorrelation tempogram extraction using librosa library [23]. ToMe attention is introduced between multi-head attention and MLP to incorporate token merging.

3 Experiments and Results

3.1 Datasets

Distress Analysis Interview Corpus/Wizard-of-Oz (DAIC-WoZ) dataset [13] is a public depression dataset available in English that contains recordings and transcripts of depressed and non-depressed individuals with 189 sessions in total, each labeled with a PHQ-8 score [16]. It consists of 30 depressed and 77 non-depressed speakers with 5068 training samples and 2679 test samples. The total duration of this dataset is roughly 14.5 h. Since the training and testing split has already been specified in the dataset, we have opted not to conduct any k-fold cross-validation during experimentation. The study also employs a Chinese dataset known as Emotional Audio-Textual Depression corpus (EATD-corpus), which consists of speech recordings from individuals classified as either depressed or non-depressed [29]. The dataset includes a total of 162 different speakers, with 30 classified as depressed and 132 as non-depressed. The categorization is based on the indexed Self-Rating Depression Scale (SDS) score (Raw SDS score multiplied by 1.25). The total duration of the dataset is around 2.26 h. Each individual provided three responses, leading to a total of 57 depressed samples and 192 non-depressed samples. Likewise, each individual in the testing set also provided three responses, resulting in a total of 33 depressed samples and 204 non-depressed samples. The exact details of both the datasets are presented in Table 2.

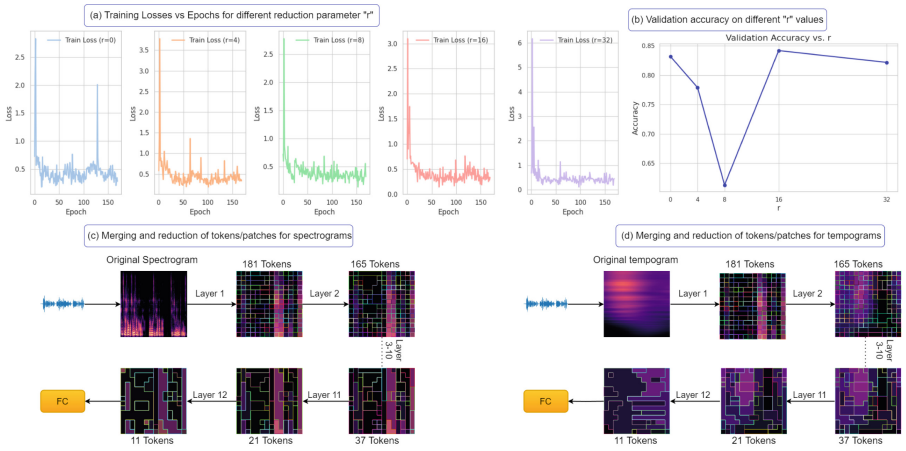


Fig. 3. Token Merging and Reduction. (a) Training Loss at different 'r' values. (b) With r = 16, validation accuracy of model is maximum around 84.3%. (c) With ToMe at r = 16, similar patches are merged in each transformer block, 16 tokens will be reduced per layer and after last layer 11 token will be left from spectrogram channel (d) 11 token from tempogram channel and fed to fully connected layer(FC) for classification.

3.2 Experimental Setup

The Pytorch [26] framework is used to implement the model. The EATD-Corpus and DAIC-WoZ dataset display inherent class imbalance, featuring fewer depressed samples compared to non-depressed samples in both the training and test sets. To mitigate this imbalance, we employed a data augmentation technique by generating multi-resolution spectrograms [10] and tempograms for each speech utterance with varying window sizes. At input layer of AST, for the depressed class in the training set of EATD-Corpus, six distinct log-mel spectrograms and tempograms were extracted for each speech utterance, using window sizes of 25 ms, 75 ms, 100 ms, 400 ms, 600 ms, and 800 ms, effectively augmenting the sample count by a factor of six. Similarly, for the depressed category in DAIC-WoZ, four window sizes of 25 ms, 100 ms, 400 ms and 800 ms were used, quadrupling the total sample count. The GPU used is Nvidia Tesla P100, the optimizer is Adam, the learning rate is kept as 3×10^{-4} with a batch size of 32. During training, early stopping epoch mechanism is used, with maximum number of epochs 50 and waiting patience 3. In the end, the average of prediction results of all features of the subject is regarded as final depression score. The F1-score, recall and precision are used to evaluate the performance.

3.3 Comparison of Training Time and Throughput

To evaluate the performance of the model on the DAIC-WoZ dataset, a comparison of training time and throughput between the baseline benchmark and

the proposed model is conducted, which is shown in Table 1. 50 runs were performed to ensure statistical robustness. The batch size was set to 32, and the input size was determined based on the default configuration of the model. The model demonstrated a significant improvement in throughput, achieving nearly a twofold increase compared to the baseline benchmark and reduced training time.

Table 1. Comparison of Training Time and Throughput.

Metric	Baseline	Proposed
Throughput	170.49 im/s	316.85 im/s
Throughput Improvement	–	1.86×
Training Time	6.52 h	3.82 h

3.4 Results and Discussion

Performance Evaluation on DAIC-WoZ Dataset: A comprehensive analysis of various methods for depression classification using audio features was conducted, focusing not only on their overall performance but also on their internal architecture contributions. As can be seen from Table 3, among the evaluated methods, the Gated Recurrent Units (GRUs) Model [29] introduced in 2022 stood out with its remarkable F1 score of 0.77 and perfect recall (1.00), indicating its robust capability to accurately detect depression. The internal architecture of the GRU Model, leveraging GRUs, likely facilitated the effective capture of temporal dependencies in the speech data, leading to superior performance. Al Hanai and others [2] demonstrated notable precision of 1.00, suggesting a high proportion of true positive predictions among all positive predictions. Additionally, the proposed ToMe models, with AST and DeiT encoders, exhibited competitive performance, with F1 scores of 0.74 and 0.77, respectively. These models introduced novel ToMe mechanisms and dual-channel architectures, enabling the effective integration of spectrograms and autocorrelation tempograms, along with attention mechanisms tailored for multimodal data fusion, thereby enhancing their ability to capture complex patterns indicative of depression across different modalities.

Performance Evaluation on EATD-Corpus Dataset: The results for the EATD-corpus is presented in Table 4. The Multi-modal LSTM [15] model achieved an F1 score of 0.49, indicating moderate performance in capturing the balance between precision and recall. These traditional machine learning models employed feature engineering techniques tailored to the characteristics of the speech data, enabling them to discern patterns indicative of depression. Furthermore, the GRU Model [29] outperformed the aforementioned methods with an

Table 2. Datasets Overview.

Dataset	DAIC-WoZ	EATD-Corpus
Language	English	Chinese
Depressed Speakers	30	30
Non-Depressed Speakers	77	132
Duration (Hours)	14.5	2.26
Training Samples	5068	249
Test Samples	2679	233
Total	7747	482

F1 score of 0.66, showcasing its ability to effectively capture temporal dependencies in the audio sequences using GRU's. The BiLSTM+Attention model [16] also demonstrated competitive performance with an F1 score of 0.65, leveraging BiLSTM units enhanced with attention mechanisms for feature representation. The proposed models, exhibited substantial improvements in performance, with F1 scores of 0.87 and 0.90, respectively.

Table 3. Results of Experiments on DAIC-WoZ dataset.

Feature	Method	F1 score	Recall	Precision
Audio	DepAudioNet [22]	0.52	1.00	0.35
	Multi-modal LSTM [15]	0.63	0.56	0.71
	GRU Model [29]	**0.77**	**1.00**	0.63
	Valtar et al. [31]	0.46	0.86	0.32
	Al Hanai et al. [2]	0.67	0.50	**1.00**
	Wei et al. [33]	0.61	0.66	0.59
	Lam et al. [19]	0.56	0.78	0.44
	BiLSTM+Attention [16]	0.73	0.72	0.78
Audio	Proposed (AST)	0.74	0.74	0.74
	Proposed (AST-DeiT)	**0.77**	0.77	0.78

Table 4. Results of Experiments on EATD-Corpus.

Feature	Method	F1 score	Recall	Precision
Audio	Multi-modal LSTM [15]	0.49	0.56	0.44
	SVM [29]	0.49	0.41	0.54
	RF [29]	0.50	0.53	0.48
	Decision Tree [29]	0.45	0.44	0.47
	GRU Model [29]	0.66	0.78	0.57
	BiLSTM+Attention [16]	0.65	0.60	0.70
Audio	Proposed (AST)	0.87	0.87	0.89
	Proposed (AST-DeiT)	**0.90**	**0.90**	**0.91**

4 Conclusion and Future Work

This paper presented a novel depression classification approach leveraging multimodal token merging within a speech spectrtemporal transformer framework. Introducing the ToMe models and a multimodal dual-channel architecture yielded promising results, particularly with ToMe dual-channel AST and ToMe dual-channel AST with DeiT encoders. Out of the two models presented, the ToMe dual-channel AST with DeiT encoders performs better, achieving an F1-score and recall of 0.90 and a precision of 0.91 for the EATD-Corpus. The model also performs better on the DAIC-WoZ dataset, achieving an F1-score and recall of 0.77 and a precision of 0.78. The study proves to be effective in outperforming all existing models for the EATD corpus and surpassing most models for the DAIC-WoZ dataset.

In future, other token pruning mechanisms based on similar representations of speech along with more advanced augmentation techniques such as DeepSMOTE [6] and SpecAugment [25] can be incorporated within the proposed framework for obtaining potential improvements in the overall performance for similar tasks. With the availability of datasets in different languages, it would be interesting to perform cross-language depression classification to build a more robust, real-world classification system that is independent of language.

References

1. Al Hanai, T., Ghassemi, M.M., Glass, J.R.: Detecting depression with audio/text sequence modeling of interviews. In: Interspeech, pp. 1716–1720 (2018)
2. Alhanai, T., Ghassemi, M., Glass, J.: Detecting depression with audio/text sequence modeling of interviews. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2018-September, pp. 1716–1720 (2018). https://doi.org/10.21437/Interspeech.2018-2522
3. Association, A.P., et al.: Diagnostic and statistical manual of mental disorders. Text revision (2000)

4. Bolya, D., Fu, C.Y., Dai, X., Zhang, P., Feichtenhofer, C., Hoffman, J.: Token merging: your vit but faster (2023)
5. Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., Quatieri, T.F.: A review of depression and suicide risk assessment using speech analysis. Speech Commun. **71**, 10–49 (2015)
6. Dablain, D., Krawczyk, B., Chawla, N.V.: Deepsmote: fusing deep learning and smote for imbalanced data. IEEE Trans. Neural Netw. Learn. Syst. **34**(9), 6390–6404 (2022)
7. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale (2021)
8. Elkin, I., et al.: National institute of mental health treatment of depression collaborative research program: general effectiveness of treatments. Arch. General Psychiat. **46**(11), 971–982 (1989). https://doi.org/10.1001/archpsyc.1989.01810110013002
9. France, D.J., Shiavi, R.G., Silverman, S., Silverman, M., Wilkes, M.: Acoustical properties of speech as indicators of depression and suicidal risk. IEEE Trans. Biomed. Eng. **47**(7), 829–837 (2000)
10. Fraser, G., Boashash, B.: Multiple window spectrogram and time-frequency distributions. In: Proceedings of ICASSP 1994. IEEE International Conference on Acoustics, Speech and Signal Processing. vol. iv, pp. IV/293–IV/296 (1994). https://doi.org/10.1109/ICASSP.1994.389818
11. Fried, E.I., Nesse, R.M.: Depression sum-scores don't add up: why analyzing specific depression symptoms is essential. BMC Med. **13**(1), 1–11 (2015)
12. Gong, Y., Chung, Y.A., Glass, J.: Ast: audio spectrogram transformer (2021)
13. Gratch, J., et al.: The distress analysis interview corpus of human and computer interviews. In: Calzolari, N. (eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pp. 3123–3128. European Language Resources Association (ELRA), Reykjavik (2014). http://www.lrec-conf.org/proceedings/lrec2014/pdf/508Paper.pdf
14. Grosche, P., Müller, M., Kurth, F.: Cyclic tempogram-a mid-level tempo representation for musicsignals (2010). https://doi.org/10.1109/ICASSP.2010.5495219
15. Hanai, T., Ghassemi, M., Glass, J.: Detecting depression with audio/text sequence modeling of interviews, pp. 1716–1720 (2018). https://doi.org/10.21437/Interspeech.2018-2522
16. Iyortsuun, N.K., Kim, S.H., Yang, H.J., Kim, S.W., Jhon, M.: Additive cross-modal attention network (acma) for depression detection based on audio and textual features. IEEE Access **12**, 20479–20489 (2024). https://doi.org/10.1109/ACCESS.2024.3362233
17. Jiang, H., et al.: Detecting depression using an ensemble logistic regression model based on multiple speech features. Comput. Math. Methods Med. **2018** (2018)
18. Kaustubh, K., Gogoi, P., Prasanna, S.: Rhythm formant analysis for automatic depression classification. In: International Conference on Speech and Computer, pp. 94–106. Springer, Heidelberg (2023). https://doi.org/10.1007/978-3-031-48309-7_8
19. Lam, G., Huang, D., Lin, W.: Context-aware deep learning for multi-modal depression detection. In: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3946–3950 (2019). https://api.semanticscholar.org/CorpusID:145833193
20. Lu, J., Liu, B., Lian, Z., Cai, C., Tao, J., Zhao, Z.: Prediction of depression severity based on transformer encoder and cnn model. In: 2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP), pp. 339–343. IEEE (2022)

21. Ma, X., Yang, H., Chen, Q., Huang, D., Wang, Y.: Depaudionet: an efficient deep model for audio based depression classification. In: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, pp. 35–42 (2016)
22. Ma, X., Yang, H., Chen, Q., Huang, D., Wang, Y.: Depaudionet: an efficient deep model for audio based depression classification. In: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge (2016). https://api.semanticscholar.org/CorpusID:2518379
23. McFee, B., et al.: librosa: audio and music signal analysis in python. In: Proceedings of the 14th Python in Science Conference, vol. 8 (2015)
24. Othmani, A., Kadoch, D., Bentounes, K., Rejaibi, E., Alfred, R., Hadid, A.: Towards robust deep neural networks for affect and depression recognition from speech. In: Del Bimbo, A., et al. (eds.) ICPR 2021. LNCS, vol. 12662, pp. 5–19. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-68790-8_1
25. Park, D.S., et al.: Specaugment: a simple data augmentation method for automatic speech recognition. arXiv preprint arXiv:1904.08779 (2019)
26. Paszke, A., et al.: Pytorch: an imperative style, high-performance deep learning library (2019). https://arxiv.org/abs/1912.01703
27. Sardari, S., Nakisa, B., Rastgoo, M.N., Eklund, P.: Audio based depression detection using convolutional autoencoder. Expert Syst. Appl. **189**, 116076 (2022)
28. Satt, A., Rozenberg, S., Hoory, R., et al.: Efficient emotion recognition from speech using deep learning on spectrograms. In: Interspeech, pp. 1089–1093 (2017)
29. Shen, Y., Yang, H., Lin, L.: Automatic depression detection: an emotional audio-textual corpus and a gru/bilstm-based model (2022)
30. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention (2021)
31. Valstar, M., et al.: Avec 2016 - depression, mood, and emotion recognition workshop and challenge (2016)
32. Verde, L., et al.: A lightweight machine learning approach to detect depression from speech analysis. In: 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI), pp. 330–335. IEEE (2021)
33. Wei, P.C., Peng, K., Roitberg, A., Yang, K., Zhang, J., Stiefelhagen, R.: Multimodal depression estimation based on sub-attentional fusion (2022)
34. Wu, P., Wang, R., Lin, H., Zhang, F., Tu, J., Sun, M.: Automatic depression recognition by intelligent speech signal processing: a systematic survey. CAAI Trans. Intell. Technol. **8**(3), 701–711 (2023)
35. Zhao, Z., et al.: Automatic assessment of depression from speech via a hierarchical attention transfer network and attention autoencoders. IEEE J. Sel. Topics Signal Process. **14**(2), 423–434 (2019)

Detecting Depression from Audio Data

Mary Idamkina[1] and Andrea Corradini[2(✉)]

[1] University of Liverpool, Liverpool, UK
maryidamkina@liverpool.ac.uk
[2] MCI, Innsbruck, Austria
andrea.corradini@mci.edu

Abstract. AI has recently started to gain popularity in most spheres of our daily life, including healthcare. While most AI-based solutions have been proposed for physical-related medical conditions, only few of them target mental health. AI-based solutions can support the diagnosis of mental disorders from clues that a human doctor may not be able to recognize, for example the biomarkers of depression. Previous studies in the area of automatic detection of depression from different media, like audio-visual data, ECG, or transcribed speech exists. In this research, we focus on the detection of depression from audio data. We use only low-level audio features and their physical characteristics, while we do not include any semantic information carried by the speech. We performed a qualitative analysis on the selected set of features that proved to be important for recognizing depression. Such an analysis revealed the dependence between gender and the set of relevant features for depression detection. We discovered differences and similarities between sets of strong predictors between gender. Formants have shown to be important to describe the articulation level for both male and female voices. Control over the voice is a better predictor for male voices, while monotony is better described with formants for male voices and with MFCC and energy for female ones. This highlights the importance of gender inclusivity in mental healthcare and depression diagnostics in particular. Experiments carried out while running this study achieved a value of F1 of 65% for depression detection for female voices and 67% for male voices.

Keywords: Machine Learning · Speech Analysis · Depression Detection

1 Introduction

Depression is a mental disorder affecting 280 million people worldwide (World Health Organization, 2023). Depression tremendously decreases the quality of life and if left untreated can even lead to a lethal ending. It is considered comparably disabling to chronic disorders such as arthritis and asthma, but with lifetime risk being 30% for men and 40% for women (Andrews & Titov, 2007). The widespread and serious consequences of depression make it crucial to learn to correctly diagnose it and track its severity over the period of treatment.

A. Karpov and V. Delić (Eds.): SPECOM 2024, LNAI 15299, pp. 336–351, 2025.
https://doi.org/10.1007/978-3-031-77961-9_25

Current diagnostics methods are mostly based on questionnaires, i.e. self-reported symptoms. While the patients themselves are the only source of truth about their thinking, experiencing symptoms is very personal and recognising them for the first time can be very challenging for correct reporting. Not surprisingly, depression is usually massively misdiagnosed by primary healthcare providers, with the rate of missed cases reaching 50% (Gómez-Gómez et al., 2022).

Depression causes changes in the brain that are reported to be seen on EEG (Zang et al., 2022), meaning there are physical changes to the brain that could indicate the presence of the condition and be used for unbiased diagnostics. Physical markers, if informative enough, would be a more reliable source of information than the reported symptoms and would be easier to track and compare reliably over time.

Depression has been shown to affect the speech centre of the brain, which makes speech a strong biomarker of Depression (Almaghrabi et al., 2023). There are multiple studies that develop Machine Learning (ML) models capable of detecting depression from speech characteristics. At the same time, recording speech is very accessible nowadays with the ubiquity of mobile devices. A tool for recognising depression from speech would mean a significant step towards delivering mental health help at low cost and without the need for human professionals or special equipment.

This study aims at investigating the properties of a human's voice and the predictive power of those properties towards diagnostics of depression. We concentrate on low-level properties and features selection, as those can be helpful for understanding the actual changes that depression causes in the patient's brain. For this research we used a publicly available dataset containing records of patients' interviews and their depression status. We extracted a set of low-level audio features from the audio records of those patients' interviews, and then we run a series of experiments on different ML models trained on this dataset. The results have been analysed, as well as the intermediary discoveries about features' relative importance and gender importance for detection of depression. The most performing ML model achieved a 65% F1 score with female voices only and a F1 score of 67% with male voices only.

2 Related Work

AI has shown impressive potential in aiding healthcare systems, including mental health, by increasing accuracy of diagnostics, optimizing costs, improving patients' education and trust towards the healthcare provider (Alowais et al, 2023).

In case of depression detection, this potential can be further exploited by using the features that would be challenging to detect for a human doctor. Traditionally depression is diagnosed by asking a patient to file a questionnaire and then analyzing their answers against a predefined scale (Tolentino and Schmidt, 2018). There are no known physical tests for depression detection. Depression has been shown to affect the speech center of the brain, which makes speech a strong biomarker of depression (Almaghrabi et al., 2023). Especially changes in the voice have proven to be strong biomarkers and can be objectively measured by extracting low-level audio features.

Speech characteristics have been consistently and successfully modeled with different Machine Learning (ML) algorithms for automatic depression detection.

An automated detection of voice changes that people with depression develop is detailed in a study where participants were asked to take a questionnaire (PHQ-9) and record two short speech samples (Zhang, 2020). The authors analyzed low-level acoustic features such as sound formants, pitch, pause-to-speech ratio etc. from 535 collected audio files and employed a Gradient Boosting tree ML method to predict depression from those features. The research concluded that prosody features are not strongly predictive of psychomotor disturbances, while some audio features like e.g. spectral slope, spectral flux, unvoiced speech segment length, loudness, MFCC's 1 and 3, mean pause length, pause variability, and pause-to speech ratio are.

Another related study that uses voice as a biomarker for depression detection is presented in (Shin, 2021). The authors recruited 93 participants and classified them to not depressed, minor depressive episode and major depressive episode. The participants were asked to answer a questionnaire to evaluate their mental state. Later, a 30 to 50-min interview with them was recorded. Voice recordings were processed to extract voice features, tempo-spectral and acoustic features. Then four Machine Learning methods were applied to classify the records. A multi-layer perceptron is the model that produced the best results obtaining an AUC of 65.9%, a sensitivity of 65.6%, and specificity of 66.2%. The authors extracted seven features showing statistical significance: spectral centroid, spectral roll-off, formant BW2, sq mean pitch, standard deviation pitch, ZCR and voice portion.

A gender dependent vowel level formant analysis in context of depression detection was performed in (Cummins, 2017). Using the DAIC-WOZ dataset, they extracted vowel level formants information from the audio data, combined those features with information about gender of the participant, and trained a linear classifier to predict depression. The maximum F1 score achieved during this study was 63%. The research again confirms that there are detectable changes in speech in case of depression. Its uniqueness is in adding a gender variable to the input, highlighting the importance of gender specific analysis of mental health issues. They also highlighted the low number of datapoints in the dataset, which drove the choice of cross-validation as means of evaluating the model performance. The choice of a linear model is also remarkable as it is more interpretable than more sophisticated ML models such as artificial neural networks and deep learning techniques in general. These findings seem to indicate that there is a benefit of using simpler models if the goal is to find real correlations between features and depression (as opposed to high performance of the resulting model).

3 Methodology and Evaluation Criteria

Depression affects the brain in general, and in particular the speech centers. That means that the audio wave the patients produce with their vocal tract can have different characteristics from those of unaffected people. Training an ML model on low-level audio features can help discover those dependencies: if depression can be detected from a change in voice, that means that this change is most probably related to depression.

3.1 Data Preparation

The DAIC-WOZ database (Gratch et al, 2014) was used for this research. It consists of 189 recorded interviews with participants of different depression statuses. The dataset contains audio recordings with transcripts, some pre-extracted features and demographical data (gender). For each participant, the response to the depression diagnosis questionnaire (PHQ-8) is provided, together with a summary score and a binary score (a thresholded summary score) used as a target variable to signify the depression status.

The dataset is split into train, dev and test parts, with 107, 35, and 47 participants respectively. The training set is used for model development, dev for model selection (including between approaches). As we did not sign a EULA, we did not obtain test set labels from the dataset owners. Due to the lack of target labels in the test set, we use dev in the testing phase for reporting final scores. Out of 142 participants in train and dev tests combined there were 63 female subjects and 79 male subjects, respectively. Out of the 63 female participants, 24 ($\approx 38\%$) had positive depression status. Out of the 79 male participants, 18 ($\approx 23\%$) had positive depression status.

3.2 Preprocessing

Before extracting features, we split the records into 20 s fragments, with only participants' speech used (transcripts are used for that). We discarded fragments shorter than 1s, as well as the interviewer's speech. We extracted the features over 10ms sliding windows with 2.5ms overlaps, which aligns well with the values used in similar works.

Due to the different number of male and female recordings, and since the recordings themselves are of different lengths, the number of fragments is different for the genders. This resulted in 736 fragments for female individuals and 1099 for male individuals.

We carried out Z-Score data normalization for computational stability. The application of normalization is used as hyperparameter, i.e. it was either applied or not applied depending on settings and in combination with other hyperparameters.

Eventually, we calculated the F1 score to estimate the model prediction quality. Precision and recall contribute equally to the target F1 score. One or another can be prioritized by applying different confidence thresholds at inference time. In this work, recall is prioritized over precision, while targeting the highest F1 score during model selection. We used the hyperparameter scale_pos_weight to determine the priority. This hyperparameter indicates the relative weight of positive examples. As the F1 score is calculated as a harmonic mean of precision and recall, in case of equals score, we select the model with higher recall according to the hyperparameter value.

3.3 Feature Extraction

Low-level audio features are extracted using two publicly available Python libraries: PyAudioAnalysis (Giannakopoulos, 2015) and OpenSmile (Eyben, Wöllmer and Schuller, 2010). PyAudioAnalysis extracts features that are used for depression classification based on audio-visual features (see Table 1). OpenSmile is capable of extracting several sets of features. We only selected OpenSmile's "eGeMAPS" and OpenSmile's

"ComParE" feature sets that include formants, loudness, MFCC, F0 (fundamental frequency) and other features. Due to an overlap between sets, we removed highly correlated and/or redundant features.

For each of the features, we then calculated these additional statistical features: the mean, standard deviation, as well as the 5th and the 95th percentile to represent soft minimum and maximum while also avoiding outliers. Beside the initial 238 values per time window, this results in 954 additional features.

4 Models

4.1 Feature Selection

Feature selection is important not only to reduce the required computational power and simplify the model, but also to avoid the "curse of dimensionality" that leads to overfitting (Li et al., 2017). With more features being used, data becomes sparse in the new feature space, and more datapoints are needed to develop a reliable model. Hence, our strategy behind our feature selection was to pick a subset of features that have the highest predictive power, but at the same time to exclude the ones that are less valuable or are redundant, meaning that can be inferred from other features.

The optimal number of features depending on the dataset size varies between classification algorithms and features correlations but it is estimated to be n-1 (where n is the number of samples i.e. the size of the dataset) for uncorrelated features and sqrt(n) for highly correlated features (Hua et al., 2004). To account for more and less conservative choices and taking into consideration sample sizes for both genders, we selected 16, 32 and 64 features and then compared the results obtained with these different feature sets. We used Recursive Feature Elimination (RFE) as a feature selection algorithm. We employed an implementation provided by the scikit-learn ML library to train a Logistic Regression model on an iteratively smaller subset of features until the desired number of features was achieved. We have carried out this procedure twice, separately for both genders.

The resulting 16 feature set for female voices and for male voices are presented in Tables 1 and 2, respectively. Feature names are abbreviated. Full names and some additional explanation of the concrete features can be found at (Opensmile, 2024).

To create the 32 feature sets for both female and male voices, we added 16 more features to the already defined 16 feature sets. This results in the two following sets listed in Tables 3 and 4, respectively.

Since no methods have shown best results per hyperparameters with 64 feature datasets, we omit to list the 64 feature sets.

4.2 Gender Considerations

Algorithmic fairness is an important consideration for ML development and has received lots of attention with the rise of AI technologies (Shrestha and Das, 2022). One of the directions of it is gender fairness: making sure that the AI model does not discriminate users based on their gender. That could happen e.g. if the data the model was trained on was biased and the model learnt to represent this bias.

Table 1. 16 feature dataset for female voice.

Feature extractor	Feature	Statistics
OpenSmile	slope500-1500_sma3, mfcc3_sma3, F1bandwidth_sma3nz, mfcc_sma_11	mean
	mfcc_sma_4	std
	audspecRasta_lengthL1norm_sma, mfcc_sma_13	p5
	HNRdBACF_sma3nz	p95
PyAudio	energy_entropy, mfcc_1, mfcc_9, mfcc_11, delta_energy_entropy	mean
	delta_spectral_centroid	std
	chroma_7	p5
	delta_chroma_9	p95

Table 2. 16 feature dataset for male voice.

Feature extractor	Feature	Statistic
OpenSmile	alphaRatio_sma3, F1frequency_sma3nz, hammarbergIndex_sma3, voicingFinalUnclipped_sma, pcm_fftMag_fband1000-4000_sma, pcm_fftMag_spectralSlope_sma	mean
	mfcc_sma_8	std
	pcm_fftMag_fband250-650_sma, pcm_fftMag_fband1000-4000_sma, pcm_fftMag_spectralHarmonicity_sma	p5
	pcm_fftMag_fband1000-4000_sma, pcm_fftMag_spectralSlope_sma, jitterLocal_sma, logHNR_sma	p95
PyAudio	mfcc_4	mean
	spectral_spread	std

The dataset we use includes gender information for each participant. This turns out to be very important to consider during the analysis because of a few findings:

- The distribution of depression statuses per gender is different both in the dataset and in reality. That means that the model would over- or under diagnose representatives of one gender, which is not inclusive.
- Men tend to have lower voices than women, so the audio features are more similar within one gender. A side effect of this difference is that is one gender has lower depression rate, it might make the model learn to predict gender instead as it would have large predictive power in that case.

Table 3. 32 feature dataset for female voice (to be added to features in Table 1).

Feature extractor	Feature	Statistic
OpenSmile	audspecRasta_lengthL1norm_sma, pcm_fftMag_spectralFlux_sma	mean
	mfcc2_sma3, mfcc_sma_9, mfcc_sma_12, mfcc_sma_13	std
	mfcc_sma_6	p5
	mfcc_sma_10, logRelF0-H1-H2_sma3nz	p95
PyAudio	mfcc_4, mfcc_6, mfcc_7, delta_energy, delta_chroma_1	mean
	delta_chroma_1	p5

Table 4. 32 feature dataset for male voice (to be added to features in Table 2).

Feature extractor	Feature	Statistic
OpenSmile	mfcc_sma_4, mfcc_sma_6, mfcc_sma_8	Mean
	pcm_fftMag_fband1000-4000_sma, mfcc_sma_2, mfcc3_sma3	Std
	F2frequency_sma3nz	p5
	F3bandwidth_sma3nz	p95
PyAudio	chroma_3	Mean
	chroma_12, delta_mfcc_9, delta_mfcc_13, delta_chroma_8	Std
	delta_spectral_flux, delta_mfcc_2	p5
	delta_spectral_entropy	p95

- During feature selection, we discovered that different features correlate more with depression status between genders. That aligns with another research (Hönig et al., 2014) that shows that spectral features are more useful for detecting male sleepiness (0.40 vs. 0.20), while prosody is more useful for detecting female sleepiness (0.33 vs. 0.21). For depression there is only a slight but similar tendency with spectral features for males 0.31 vs. for female 0.26 and prosody for female voices 0.35 vs. male voices 0.33.

The gender imbalance is not only present in automatic diagnostics but also in reporting. Previous research (Angst and Dobler-Mikola, 1984) shows that in general women tend to report more symptoms while men more likely forget symptoms, frequency and length of less recent depressions more readily. Moreover, women have a preference to see a physician or proceed to self-medication much more often than men (Angst and Dobler-Mikola, 1984), even though the reasons, being them either biological or social, are not clear. There is also a gender difference in the ways people distinguish between real and reported presence of symptoms i.e. whether men really suffer from depression

symptoms less often or tend to underreport them. If the latter is the reason, automatic diagnostics would be initially complicated at development stage because of imbalanced data, but would lead to less biased results, as the automatic feature extraction would not suffer from underreporting gender-related symptoms.

To account for that gender related difference, we split the dataset into two parts according to the participant's gender and the analysis was performed for the two genders separately. Different features were then extracted and provided to different models. From the user perspective, they will be asked to provide their gender together with the voice sample, and the program would call a corresponding ML model.

4.3 Model Selection

We considered a wide range of ML models for solving the problem. Eventually, we focus on the following three ones:

- Artificial Neural Networks (ANNs) with fully connected layers, notably Multi-layer Perceptron (MLPs), trained on statistical features
- Support Vector Machine (SVM) trained on statistical features
- XGBoost trained on statistical features

Selection between the models was performed based on the experimental results, with quality (F1) being the main criteria. Time performance was not considered as a selection criterion, for it is not targeted in this research.

Another criterion that was not considered initially but the selected model happened to possess is explainability. That characteristic is very important for interpreting the research results. Since the model is able to provide information about what features are the most important, from this information we can infer the real changes to the voice that most probably predict the depression status.

5 Results

For each model, we run a series of experiments with varying hyperparameters. Models have been trained on train set for each gender and quality assessment were measured using a dev set. We carried out hyperparameter tuning for both genders independently.

5.1 Multi-layer Perceptron

We tuned the following hyperparameters of our fully connected ANNs:

- The number of features used: 16, 32 or 64
- The number of fully connected layers
- The dimensionality of the layers
- Normalization i.e. whether to apply data normalization to the input

The hyperparameters combinations and corresponding F1 scores are represented in Table 5 for female voices and Table 6 for male voices.

This model leads to a highest achieved F1 value of 0.38 for female voices and 0.48 for male voices, respectively.

Table 5. Hyperparameters tuning of MLPs for female voices.

Features used	Layers	Dimensionality	Data normalized	Precision	Recall	F1
64	5	32	no	–	–	0.0
64	5	32	yes	0.31	0.14	0.20
32	5	32	no	0.38	0.38	0.38
32	5	32	yes	0.34	0.13	0.19
16	5	16	no	0.25	0.18	0.21
16	5	16	yes	0.31	0.18	0.23
16	3	16	no	0.20	0.11	0.14
16	3	16	yes	0.20	0.10	0.13

Table 6. Hyperparameters tuning of MLPs for male voices.

Features used	Layers	Dimensionality	Data normalized	Precision	Recall	F1
64	5	32	no	–	–	0.0
64	5	32	yes	0.13	0.02	0.03
32	5	32	no	0.13	0.04	0.06
32	5	32	yes	0.10	0.01	0.02
16	5	16	no	0.0	0.0	0.0
16	5	16	yes	0.56	0.42	0.48
16	3	16	no	0.74	0.32	0.44
16	3	16	yes	0.29	0.13	0.18
16	7	16	yes	0.51	0.45	0.48
16	7	16	yes	0.46	0.41	0.47

5.2 Support Vector Machine (SVM)

We tuned the following major hyperparameters for SVM training and classification:

- Number of features used: 16, 32 or 64.
- C: a hyperparameter controlling overfitting. Responds for the allowed margin between the classes at cost of some misclassified instances.
- Kernel: a function defining the distance between input vectors. Using different kernels allows to model non-linear dependencies.

The hyperparameters combinations and corresponding F1 scores are represented in Table 7 for female voices and Table 8 for male voices.

The highest F1 achieved was 0.26 and 0.48 for female and male voices, respectively.

Table 7. Hyperparameters tuning of SVM for female voices.

No. of features	C	Kernel	Precision	Recall	F1
16	1.0	linear	0.3	0.23	0.26
32	1.0	linear	0.31	0.19	0.23
64	1.0	linear	0.27	0.21	0.23
16	1.0	rbf	0.0	0.0	0.0
32	1.0	rbf	0.0	0.0	0.0
64	1.0	rbf	0.0	0.0	0.0
16	1.0	sigmoid	0.0	0.0	0.0
32	1.0	sigmoid	0.0	0.0	0.0
64	1.0	sigmoid	0.0	0.0	0.0
16	0.1	linear	0.28	0.19	0.22
16	0.5	linear	0.31	0.22	0.26
16	0.25	linear	0.30	0.20	0.24

Table 8. Hyperparameters tuning of SVM for male voices.

No. of features	C	Kernel	Precision	Recall	F1
16	1.0	linear	0.45	0.51	0.48
32	1.0	linear	0.28	0.31	0.29
16	1.0	linear	0.35	0.42	0.38
32	1.0	linear	0.52	0.42	0.46
64	1.0	linear	–	–	no converge
16	0.1	linear	0.41	0.44	0.42
16	0.5	linear	0.46	0.50	0.48
32	0.5	linear	0.33	0.29	0.30

5.3 Xgboost

We tuned the following hyperparameters of the XGBoost model:

- Number of features: 32 or 64. In case of XGBoost, 32 features subset is the first 32 features of the 64 features set, which gave the best results. They are all from the OpenSmile feature sets.
- Number of estimators: the number of decision trees in the ensemble.
- Maximum depth: the maximum depth of each decision tree in the ensemble.
- ETA or learning rate: weight of the new trees added to the ensemble. Setting it to a lower value helps decrease overfitting.
- Gamma: minimum loss reduction for splitting the decision tree.

- Subsample: ratio of the dataset that is used for building each new tree.
- Scale_pos_weight: the relative weight of the positive examples. Setting it to a higher value means prioritizing Recall over Precision.

The hyperparameters combinations and corresponding F1 scores are represented in Table 9 for female voices and Table 10 for male voices.

Table 9. Hyperparameters tuning of XGBoost for female voices.

No. of features	estima-tors	max_depth	eta	gamma	sub-sample	scale_pos_weight	Precision	Recall	F1
32	100	5	0.3	0	1	2.5	0.29	0.24	0.26
32	100	3	0.3	0	1	2.5	0.27	0.17	0.20
32	200	5	0.3	0	1	2.5	0.30	0.24	0.27
32	200	5	1e−3	0	1	2.5	0.43	0.43	0.43
32	200	5	1e−3	10	1	2.5	0.43	0.44	0.43
32	200	5	1e−3	10	0.5	2.5	0.36	0.36	0.36
32	200	5	1e−3	10	0.75	2.5	0.4	0.4	0.40
32	200	4	1e−3	10	1	2.5	0.44	0.53	0.48
32	200	3	1e−3	10	1	2.5	0.47	0.58	0.52
32	100	3	1e−3	10	1	2.5	0.46	0.62	0.52
32	100	3	1e−3	10	1	3.5	0.45	0.70	0.54
32	100	3	1e−4	10	1	3.5	0.46	0.73	0.57
32	100	2	1e−4	10	1	3.5	0.49	0.94	0.65
32	100	2	1e−4	10	1	2.5	0.50	0.9	0.65
64	100	5	0.3	0	1	2.5	0.34	0.22	0.27
64	100	3	0.3	0	1	2.5	0.31	0.20	0.24
64	100	3	1e−2	0	1	2.5	0.42	0.44	0.43
64	50	3	1e−2	0	1	2.5	0.48	0.54	0.51
64	50	3	1e−3	0	1	2.5	0.45	0.61	0.52
64	50	3	1e−3	0	1	3.5	0.45	0.61	0.52
64	50	2	1e−3	0	1	2.5	0.42	0.58	0.49
64	100	2	1e−3	0	1	2.5	0.42	0.58	0.49
64	100	2	1e−3	0	0.5	2.5	0.45	0.4	0.42
64	100	2	1e−3	10	1	2.5	0.42	0.58	0.49

As it can be evinced from Tables 9 and 10, the highest F1 achieved was 0.65 for female voices and 0.67 for male ones. In case of female voices, the 32 features used leading to the highest F1 score were all from OpenSmile and were the following ones: alphaRatio_sma3, slope500-1500_sma3, mfcc2_sma3, mfcc3_sma3, mfcc4_sma3, mfcc4_sma3, F0semitoneFrom27.5Hz_sma3nz, HNRdBACF_sma3nz, logRelF0-H1-H2_sma3nz, jitterDDP_sma, mfcc_sma_9_, logRelF0-H1-A3_sma3nz, F2frequency_sma3nz, pcm_fftMag_fband1000-4000_sma, shimmerLocaldB_sma3nz,

F1bandwidth_sma3nz, mfcc_sma_4_, audspecRasta_lengthL1norm_sma, pcm_fftMag_spectralVariance_sma,mfcc_sma_3_, pcm_fftMag_spectralFlux_sma, shimmerLocal_sma, jitterDDP_sma, pcm_fftMag_spectralEntropy_sma, audspecRasta_lengthL1norm_sma,mfcc_sma_2_, pcm_fftMag_spectralKurtosis_sma, pcm_fftMag_spectralFlux_sma, mfcc_sma_6_, pcm_fftMag_spectralKurtosis_sma and pcm_fftMag_spectralRollOff25.0_sma. In case of female voices, the 32 features used leading to the highest F1 score were all from OpenSmile and were the following ones: alphaRatio_sma3, hammarbergIndex_sma3, mfcc1_sma3, mfcc3_sma3, mfcc3_sma3, mfcc4_sma3 F0semitoneFrom27.5Hz_sma3nz, HNRdBACF_sma3nz, logRelF0-H1-A3_sma3nz, logRelF0-H1-A3_sma3nz, F1frequency_sma3nz, F2frequency_sma3nz, logHNR_sma, F2bandwidth_sma3nz, F3bandwidth_sma3nz, voicingFinalUnclipped_sma, logHNR_sma, audspec_lengthL1norm_sma, audspecRasta_lengthL1norm_sma, mfcc_sma_2_, pcm_fftMag_fband250-650_sma pcm_fftMag_fband1000-4000_sma, mfcc_sma_4_, pcm_fftMag_fband1000-4000_sma, pcm_fftMag_fband1000-4000_sma, pcm_fftMag_fband1000-4000_sma, mfcc_sma_4_, jitterLocal_sma, pcm_fftMag_spectralVariance_sma, mfcc_sma_6_, pcm_fftMag_spectralSlope_sma, pcm_fftMag_spectralSlope_sma and pcm_fftMag_spectralHarmonicity_sma.

Table 10. Hyperparameters tuning of XGBoost for male voices.

No. of features	estima-tors	max_depth	eta	gamma	sub-sample	scale_pos_weight	Precision	Recall	F1
32	200	2	1e−4	0	0.5	5.5	0.57	0.79	0.67
32	200	2	1e−3	0	0.5	5.5	0.57	0.79	0.66
32	200	3	1e−3	0	0.5	5.5	0.57	0.74	0.64
32	200	3	1e−3	10	0.5	5.5	0.57	0.74	0.64
64	100	5	0.3	0	1	4.7	0.14	0.06	0.08
64	100	3	0.3	0	1	4.7	0.13	0.05	0.06
64	100	3	1e−2	0	1	4.7	0.44	0.38	0.40
64	100	3	1e−2	0	1	5.5	0.48	0.54	0.50
64	200	3	1e−2	0	1	5.5	0.42	0.42	0.42
64	200	3	1e−3	0	1	5.5	0.60	0.66	0.63
64	200	2	1e−3	0	1	5.5	0.54	0.78	0.63
64	500	2	1e−3	0	1	5.5	0.52	0.78	0.62
64	200	2	1e−3	0	1	5.5	0.51	0.79	0.62
64	200	2	1e−2	0	1	5.5	0.44	0.57	0.50
64	200	2	1e−4	0	0.5	5.5	0.56	0.79	0.66

5.4 Comparison of Results Over the Models

The best achieved F1 was shown by the XGBoost model and resulted in 65% for female voices and 67% for male voices. This is comparable to results of similar research in the

field accounting for the type of input data (audio) and only using low-level audio features. As XGBoost shows the highest F1 among all ML models, its highest explainability also makes it the best candidate for the final model to use in classification and production.

5.5 Gender Importance

Gender has been discovered to be an important factor in depression diagnostics, to a level that different features are most useful for different gender. In the smallest 16 features set, there is a zero intersection between genders. That result highlights the complex nature of mental health issues, as well as an increased need for inclusivity in research and providing care.

5.6 Feature Importance

There are features that seem to be relevant for both genders when it comes to the detection of depression and there are also gender-specific ones. Below is presented the explanation of what those features describe physically and how those align with common beliefs of depressed speech.

Features Common for Both Genders. F1 formant proved to be an important feature for depression detection for both genders (but in different statistical forms). Practically, F1 is a formant that covaries with the mouth opening and closing cycles. These cycles are the articulatory basis of speech rhythm and play a crucial role in speech comprehension (He, Zhang and Dellwo, 2019). Hence, this feature correlates with speech being better or worse articulated.

Harmonic to Noise Ratio (HNR), has also shown predictive power for both genders. HNR is a ratio between the clean voice and the sub product noise coming from the air interacting with vocal cords and other parts of the vocal tract. It can be used to evaluate the general health of the voice, e.g. detect laryngeal pathologies (Teixeira, Oliveira and Lopes, 2013).

Features More Important for Male Patients. F2 and F3 formants turned out to have crucial importance for male gender only. Those formants are dependent on articulation. F2 depends on front and back position of the tongue while F3 is important for quality and clarity of pronounced phoneme (Prica and Ilic, 2010). This aligns with the judgement of speech changes towards more incomprehensive and less articulated during depression.

Alpha Ratio and Jitter also turned out to be more important for diagnosis in male voices. Alpha ratio evaluates the relationship between low harmonics and those above 1000Hz, and it is associated with spectral slope, the delivery of information about the source as well as the filter and voice quality. Hypofunctional voices are associated with low Alpha ratio values, while hyper functional voices are associated with high Alpha Ratio values (Marsano-Cornejo and Roco-Videla, 2023). Jitter evaluates perturbation of the fundamental frequency by comparing one cycle with another (Marsano-Cornejo and Roco-Videla, 2023) and can be used to describe the voice stability. The importance of those features for depression detection in male voices may indicate how the disease affects the patient's ability to control their voice.

The Hammarberg index is the intensity difference between higher and lower frequency bands. It is also described as a measure of vocal effort and signifies the intensity of the voice, also used for emotions recognition (Schmidt, Janse and Scharenborg, 2016).

A few magnitude spectrum features related to the F0, or fundamental frequency, turned out to have predictive power of depression for male voices too. F0 variation can be thought of as what is removed by speaking in monotone (Brown and Bacon, 2010), which also aligns with the belief of speech of depressed patients being monotonous.

Features More Important for Female Patients. MFCC turned out to be mostly important for female gender. MFCC, or Mel Frequency Cepstral Coefficients, audio features that are known to describe vocal tract characteristics and are used in word separation and emotion recognition (Abdul and Al-Talabani, 2022). As a spectral feature, it can be corresponding to the speech being more monotonous in depressed patients.

Energy entropy and a delta of it proved to be relevant for the problem for female voices. This is a measure of abrupt changes in audio signal. In case of voice this might indicate whether the speech is monotonous or rather expressive, with lots of changes to the loudness.

Spectral centroid frequency (SCF) is the average frequency for a given sub-band, weighted by the normalized energy for this sub-band. This feature is affected by changes in pitch and harmonic structure and contains formants information (Kua et al., 2010).

6 Conclusion and Discussion

This research achieved F1 of 65% for depression detection for female voices and 67% for male voices. In another study (Vlasenko, 2017) that utilized an experimental setting similar to the one we employed for our research, the best F1 score of 100% was reported for female speakers. Such a result significantly exceeds the results in our current work and we could not replicate it. The difference in results is likely due to the authors using Vowel Level formant features. To extract such features, they performed a phonetic transcription of the speech and then identified the phoneme borders. In our approach, we instead calculated the formant features over the whole sound sample.

Features responsible for articulation and overall voice health proved to be important for both male and female voices. Depression in male voices could be better predicted by formant and harmonic features, responsible for monotony and ability to control the voice. For female voices, monotony also proved to be important, but is better predicted via MFCC, energy entropy and spectral centroid features.

The study also revealed the significance of gender for determining a set or features with high predictive power towards depression detection. It would be interesting to analyze the influence of other demographic characteristics in this context, for example age, level of education and marital status.

Another potential direction of further research is to investigate whether the changes in speech persist for different languages and whether the same low-level audio features would preserve predictive strength for languages other than English. While there are no reasons to doubt that depression symptoms themselves would still be present in non-English native speakers, such an analysis would help distinguish linguistic features from pure sound formation ones.

To more reliably track the changes in the voice, collecting data from different periods of the same patient's treatment would open an opportunity to distinguish personal voice characteristics from depression biomarkers, and also extract the features more reliably. For example, the speech of depressed people is often described as retarded and monotonous. While with current data the only baseline is average speaking pace, in case of the same patient it would be possible to compare whether they are really speaking slower than their normal pace or it is a personal characteristic.

While our study treated the problem as binary classification, it would be interesting to verify whether voice properties are granular enough to predict the actual score on the depression questionnaire, i.e. predict not only a binary label but also a quantitative value. The same idea can be extended and applied to predicting rubrics: e.g. instead of predicting the depression status as a whole, target only tiredness, depressed moods, and /or other specific rubrics of the depression questionnaire.

References

World Health Organization. Depressive disorder (depression) (2023). https://www.who.int/news-room/fact-sheets/detail/depression. Accessed 16 Sep 2024

Andrews, G., Titov, N.: Depression is very disabling. The Lancet **370**(9590), 808–809 (2007)

Cummins, N., Vlasenko, B., Sagha, H., Schuller, B.: Enhancing speech-based depression detection through gender dependent vowel-level formant features. In: ten Teije, A., Popow, C., Holmes, J., Sacchi, L. (eds.) Lecture Notes in Computer Science, vol. 10259, pp. 209–214. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59758-4_23

Gómez-Gómez, I., et al.: Utility of PHQ-2, PHQ-8 and PHQ-9 for detecting major depression in primary health care: a validation study in Spain. Psychol. Med. **53**, 5625–5635 (2022)

Almaghrabi, S.A., Clark, S.R., Baumert, M.: Bio-acoustic features of depression: a Review. Biomed. Signal Process. Control **85**, 105020 (2023)

Alowais, S.A., et al.: Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med. Educ. **23**, 689 (2023)

Tolentino, J.C., Schmidt, S.L.: DSM-5 criteria and depression severity: implications for clinical practice. Front. Psychiatry **9**, 450 (2018)

Gratch, J., et al: The distress analysis interview corpus of human and computer interviews. In: Proceedings of Language Resources and Evaluation Conference, pp. 3123–3128 (2014)

Eyben, F., Wöllmer, M., Schuller, B.: Opensmile. In: Proceedings of the 18th ACM International Conference on Multimedia, pp. 1459–1462 (2010)

Giannakopoulos, T.: Feature extraction (2015). https://github.com/tyiannak/pyAudioAnalysis/wiki/3.-Feature-Extraction. Accessed 16 Sep 2024

Li, J., et al.: Feature selection. ACM Comput. Surv. **50**(6), 1–45 (2017)

Hua, J., et al.: Optimal number of features as a function of sample size for various classification rules. Bioinformatics **21**(8), 1509–1515 (2004)

Opensmile: Feature set (2024). https://audeering.github.io/opensmile-python/api/opensmile.FeatureSet.html#featureset. Accessed 16 Sep 2024

Shrestha, S. and Das, S.: Exploring gender biases in ML and AI academic research through systematic literature review. Front. Artif. Intell. **5**, 976838 (2022)

Hönig, F., et al.: Automatic modelling of depressed speech: relevant features and relevance of gender. In: Proceedings of 15th Interspeech (2014)

Angst, J., Dobler-Mikola, A.: Do the diagnostic criteria determine the sex ratio in depression? J. Affect. Disord. **7**(3–4), 189–198 (1984)

He, L., Zhang, Y., Dellwo, V.: Between-speaker variability and temporal organization of the first formant. J. Acoust. Soc. Am. **145**(3), EL209 (2019)

Teixeira, J.P., Oliveira, C., Lopes, C.: Vocal acoustic analysis – jitter, shimmer and HNR parameters. Procedia Technol. **9**, 1112–1122 (2013)

Prica, B., Ilic, S.: Recognition of vowels in continuous speech by using formants. Facta universitatis - Series Electron. Energetics **23**(3), 379–393 (2010)

Marsano-Cornejo, M.-J., Roco-Videla, Á.: Variation of the acoustic parameters: F0, jitter, shimmer and alpha ratio in relation with different background noise levels. Acta Otorrinolaringologica (Engl. Ed.) **74**(4), 219–225 (2023)

Schmidt, J., Janse, E., Scharenborg, O.: Perception of emotion in conversational speech by younger and older listeners. Front. Psychol. **7**, 781 (2016)

Shin, D., et al.: Detection of minor and major depression through voice as a biomarker using machine learning. J. Clin. Med. **10**(14), 3046 (2021)

Abdul, Z.K., Al-Talabani, A.K.: Mel frequency cepstral coefficient and its applications: a review. IEEE Access **10**, 122136–122158 (2022)

Kua, J.M.K., Thiruvaran, T., Nosratighods, M., Ambikairajah, E., Epps, J.: Investigation of spectral centroid magnitude and frequency for speaker recognition. In Odyssey, p. 7 (2010)

Zhang, L., et al.: Automated voice biomarkers for depression symptoms using an online cross-sectional data collection initiative. Depress. Anxiety **37**(7), 657–669 (2020)

Vlasenko, B., Sagha, H., Cummins, N., Schuller, B.: Implementing gender-dependent vowel-level analysis for boosting speech-based depression recognition. In: Proceedings of the 18th Interspeech, pp. 3266–3270 (2017)

Zang, X., Li, B., Zhao, L., Yan, D., Yang, L.: End-to-End Depression Recognition Based on a One-Dimensional Convolution Neural Network Model Using Two-Lead ECG Signal. J. Med. Biol. Eng. **42**(2), 225–233 (2022)

Binary and Multiclass Classification of Dysphonia Using Whisper Encoder and One-Dimensional Convolutional Neural Network

Dosti Aziz[1,2(✉)] and Dávid Sztahó[1]

[1] Department of Telecommunication and Artificial Intelligence, Budapest University of Technology and Economics, Magyar tudósok körútja 2., 1117 Budapest, Hungary
azizd@edu.bme.hu, sztaho.david@vik.bme.hu
[2] Computer Science Department, University of Sulaimani, Qlyasan Street, 46001 Sulaymaniyah, Iraq

Abstract. Dysphonia is a condition characterized by difficulties in voice production caused by functional, psychological, and neurological factors. Accurate diagnosis of dysphonia is crucial for determining the proper treatment procedures and follow-ups. The previous approaches primarily focused on sustained vowels and manual feature extraction, which does not represent everyday speech usage. It raised questions regarding the usability of these methodologies when applied in real-life scenarios. Also, differentiating between organic and functional dysphonia remains an unexplored area in speech pathology. In this paper, we propose an approach based on Web-scale Supervised Pretraining for Speech Recognition (Whisper). Features extracted from the pre-trained transformer-based Whisper encoder in both Base and Large variations were used to train machine learning models such as Support Vector Machine, Random Forest, and Multi-Layer Perceptron. We also proposed an architecture based on a 1-dimensional convolution neural network (1DCNN). Our proposed method showed high performance and surpassed previous approaches on the same dataset in binary and multiclass classification. It achieved 95.51% accuracy in binary classification and 76.40% in multiclass classification, outperforming previous methods. These results emphasize the effectiveness of our model in capturing speech characteristics at the utterance level and distinguishing between dysphonia subtypes.

Keywords: Dysphonia · Whisper · Speech embeddings · Speech disorder

1 Introduction

Although human speech is the primary communication medium between individuals, it also contains essential information regarding the speaker's health status, sex, emotional state, and identity [10,20]. As speech is the outcome of a complex

A. Karpov and V. Delić (Eds.): SPECOM 2024, LNAI 15299, pp. 352–366, 2025.
https://doi.org/10.1007/978-3-031-77961-9_26

process involving different human organs, any condition that affects these organs will have an impact on the produced speech. Voice pathology can result from various factors, including organic, structural, and functional [3,33]; these conditions affect the vocal folds responsible for producing speech. Voice pathology refers to a condition in which there are abnormalities in overall voice quality, pitch, and loudness in an individual's speech compared to those of the same age, sex, and cultural background [19]. The percentage of the population affected by this condition is around 10% in the general population, but it is higher in individuals with extensive vocal fold usage, reaching almost 50% [2,16,26].

Dysphonia refers to difficulties with voice production, resulting in changes in voice quality. It is important to note that while hoarseness is a symptom reported by patients, dysphonia is a clinical diagnosis made by healthcare professionals [15]. Generally, dysphonia is categorized into two subtypes: organic dysphonia, which is the consequence of neurodegenerative diseases and structural conditions. Functional dysphonia results from the improper functioning of the vocal folds with no correlation to neurological disease [7]. These speech-related conditions can significantly affect an individual's quality of life, potentially limiting their professional performance and restricting their ability to engage with others. Therefore, proper diagnosis of these conditions is crucial.

Diagnosing dysphonia involves various procedures. The first technique is an invasive procedure that utilizes methods such as laryngoscopy, stroboscopy, and laryngeal electromyography, performed by specialized healthcare professionals and voice therapists. The main issues with these protocols are their time-consuming nature, high cost, and the discomfort they often cause patients, which is why many individuals avoid undergoing these procedures when needed [27,30]. An alternative approach involves using speech and signal-processing techniques combined with machine learning (ML) and deep learning (DL). These techniques have gained considerable attention from clinicians and academics due to their non-invasive, inexpensive, and objective nature. The methodologies typically consist of three main steps: first, collecting and acquiring speech samples; second, feature extraction, which can sometimes be bypassed when using DL techniques; and third, training the model to distinguish between normal and dysphonic speech.

Some research that falls into the ML category has primarily focused on using sustained vowels for building the detection system, including [9,13,17,18,32]. A mixture of VGG16 and a Support Vector Machine (SVM) was presented in [24]. They utilized features extracted from the sustained /*i*/ vowel using VGG16 to train the SVM classifier, achieving an accuracy of 96.7%. Other studies have investigated the effectiveness of algorithms such as K-nearest neighbors (KNN), SVM, and random forest (RF) with features extracted from vowels /*a*/, /*i*/, and /*u*/ produced at a normal, high, low pitch [8]. While highly accurate, these approaches do not reflect everyday speech usage and might downgrade performance when applied in real-life scenarios. This leads other researchers to utilize continuous speech instead of sustained vowels, which capture more variation in speech signals and represent real-life speech usage. In [29], an accuracy of

89% was achieved using continuous speech samples in binary classification. A combination of features from vowels and sentences was proposed to estimate the severity of dysphonia [34]. The issue with traditional ML approaches is that they rely on manual feature extraction, which is time-consuming and requires background knowledge to determine the best-performing features for specific speech disorders and languages.

In recent years, voice pathology detection systems have shifted to adapting deep learning (DL) networks and embedding feature features from general speech processing models. Some research has adopted Convolutional Neural Networks (CNNs), including [1,6,12,14], with accuracies ranging from 71% to 98%. A mel-spectrogram extracted from sustained vowels was used with a CNN for binary classification of dysphonia and achieved an accuracy of 92%. A combination of Bi-directional long short-term memory (biLSTM) and CNN models was proposed in [1,12] and trained using the raw waveform from a speech signal. The models achieved an accuracy of 98.6% and 71.36%. Other researchers used embedding features from self-supervised models, such as in [11] for Parkinson's detection and [25], which utilized self-supervised models for dysphonia detection, achieving the highest accuracy of 94% in binary classification.

Despite the limited research available on differentiating between functional and organic dysphonia using speech samples combined with ML and DL, it is crucial to make this distinction due to the different treatment procedures required for each dysphonia category. Functional dysphonia may require voice therapy or psychological intervention, while organic dysphonia may necessitate medical intervention or surgery. The similarities in effects these subtypes have on speech signals may explain the challenges in distinguishing between them. To our knowledge, only [4,29] have performed classifications between organic and functional dysphonia, achieving accuracies of 75% and 73.26%, respectively.

Existing studies in the literature have several limitations. They mainly focused on sustained vowels and manual feature extraction, which is expensive, time-consuming, and requires background knowledge. Moreover, using sustained vowels neither captures all speech variation nor reflects everyday speech usage, which limits their usability in real-world applications. DL approaches require large datasets, which are not available in the speech pathology domain due to the data and patient privacy, and this leads these methodologies to not perform well in detecting voice pathology. Also, the differentiating between functional and organic dysphonia remains the primary unexplored area.

In this paper, we propose a non-invasive method that eliminates the manual feature extraction method to enhance the classification performance of dysphonia diagnosis in both binary and multiclass classification. Our approach is based on the embedding feature from the transformer encoder part of the Web-scale Supervised Pretraining for Speech Recognition (Whisper) model. This transformer-based speech processing model is able to capture various acoustic and speaker-related features in a non-invasive manner and outperform conventional acoustic features in distinguishing between normal and dysphonic speech. Our contributions to the paper are as follows.

- Adopt Whisper encoder in both Base and Large variations using a transfer learning approach for diagnosing dysphonic speech.
- Propose a custom 1DCNN on top of the Whisper transformer encoder for both binary and multiclass dysphonia classification.
- Our proposed approach outperformed previous related work in the literature in both binary and multiclass cases.

The remaining sections of this paper are as follows. In Sect. 2, we include a detailed explanation of the dataset used in the experiment and the architecture of our custom 1DCNN model. Section 3 will explain the results achieved, followed by the discussion section in Sect. 4. The conclusion and future work will be described in Sect. 5.

2 Materials and Method

2.1 Datasets

The study utilized a Hungarian dataset containing speech samples from both healthy control (HC) and those with organic (OD) and functional dysphonia (FD) groups. Native Hungarian speakers were tasked with reading a translated tale, "The North Wind and the Sun," in a clinical office at the Head and Neck Surgery Department of the National Institute of Oncology. Three Experts in the same department categorized patients into organic and functional dysphonia, and labels were determined by the average of three. The recordings were in PCM audio format with a 16 and 44 kHz sampling rate and 16-bit quantization, and the duration of one speech sample was about 45 seconds. Permission was obtained for the use of speech samples in research.

The dataset consists of a total of 441 speech samples, including 178 samples from individuals with OD (98 females and 81 males), 83 FD (62 females and 21 males), and 178 from HC (93 females and 86 males). Although the original sampling rates varied, all speech samples were resampled to a 16 kHz rate to align with the Whisper feature encoder used in the experiments. Figure 1 shows each class's distribution of male and female samples.

The dataset was separated into training and testing sets, ensuring representative samples for each class were included in the test set. The ratio between the training and testing sets was 80% and 20%, respectively. The 80% of the training set was used for 5-fold cross-validation for hyperparameter and model selection.

2.2 Whisper Model

Whisper is an encoder-decoder model based on the standard transformer architecture in [31]. The model was trained using weak supervision on a large-scale dataset, encompassing up to 680,000 hours of multilingual data, and trained using a multitasking approach, including speech recognition, speech activity detection, speaker diarization, and more. Log-magnitude mel-spectrograms were

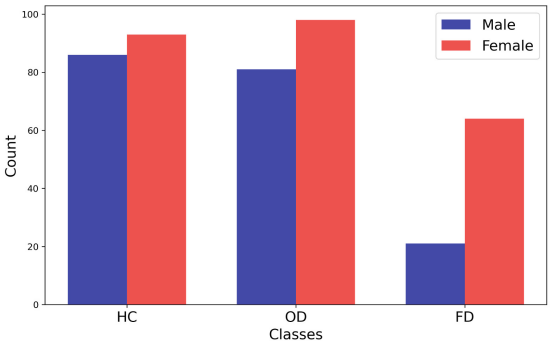


Fig. 1. Sex Distribution of Each Class in the Dataset.

extracted using a 25-millisecond window and a 10-millisecond overlap, and the training utilized the AdamW optimizer with a linear learning rate decay, gradient norm clipping, and a batch size of 256. Whisper models come in various sizes, reflecting the number of transformer layers utilized. The Tiny model, the smallest, contains 4 transformer layers. The Base model includes 6 layers, and the Large model consists of 32 transformer layers [22]. In our experiment, we adopted the Base and Large models, which have 39 million and 1550 million parameters, respectively.

2.3 Proposed Method

This paper utilized two variations of the Whisper encoder, Base and Large model. The pre-trained versions of two models were utilized using the SpeechBrain toolkit [23] and the Transformers library from Hugging Face. These models can process variable-length speech signals and convert them into fixed-length speech embeddings. The Base and Large models have dimensions of 512 and 1280, respectively.

The model will extract the log mel-spectrogram from each sample, which will serve as the input for the encoder part of the model. During this process, the parameters of the encoder layers are kept frozen. This technique involves transferring knowledge from one domain to different domains and datasets, a method known as transfer learning. Transfer learning has shown promising results, especially when the dataset size is insufficient for training the model from scratch or fine-tuning the model. The output of the encoder section will be a matrix with a shape of $1 \times 1500 \times \text{Dim}$, indicating the batch size, timestep, and feature dimensionality, respectively. After applying average pooling along the timestep dimension, we get a feature vector of shape $1 \times \text{Dim}$, representing a feature vector capturing speech characteristics relevant to the overall speech sample, whether healthy or dysphonic.

In this study, we used this feature vector to train various ML algorithms, namely SVM with both rbf and linear kernels, RF, and MLP. Additionally, we

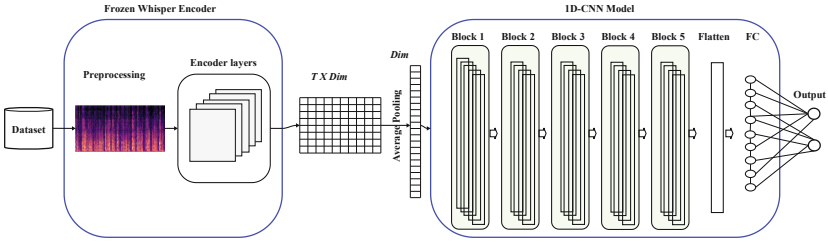


Fig. 2. Architecture of the proposed method.

proposed an architecture based on a 1-dimensional convolutional neural network (1DCNN). Figure 2 illustrates the architecture of the proposed method.

The CNN architecture consists of five CNN blocks. The first block has a kernel size of 5, an input channel of 1 (reflecting the single channel in our input feature vector), and 64 output channels. The convolution layers are followed by batch normalization, max-pooling with a kernel size of 2 and a stride of 2, and a dropout layer to prevent overfitting. The remaining four blocks follow the same structure but with a kernel size of 3 and an increasing number of channels: 128, 256, 512, and 1024, respectively. After applying adaptive global average pooling, a feature vector of size 1024 dimensions is obtained and fed to a linear layer with softmax activation function output neurons of 2 and 3, representing the two classes in the dataset for binary and multiclass classification, respectively. The PyTorch framework was used for training and development.

Finding the best hyperparameters of the ML algorithm performed using grid search and 5-fold cross-validation in scikit-learn library [21]. The best hyperparameters were used to train the ML algorithms on the full training set, and the performance of the trained model was evaluated on an independent test set. Table 1 shows the best hyperparameter obtained for binary/multiclass classification.

Table 1. List of best hyperparameter values of each algorithm.

Algorithms	Model size	Best Hyperparameters
SVM-rbf	Large	C: 2.0/3.0, gamma: 0.401/0.201
	Base	C: 1.5/9.0, gamma: 0.201/0.201
SVM-linear	Large	C: 2.0/1.5
	Base	C: 1.0/1.0
RF	Large	max-depth: 5/None, min-samples-split: 5/20, num-estimators: 100/50
	Base	max-depth: None, min-samples-split: 2/10, num-estimators: 200/500
MLP	Large	layer-sizes: (50, 50)/(200,200), max-iter: 1000/1000, Adam, lr: 0.001, batch: 4/4, loss log
	Base	layer-sizes: (200, 200)/(100,100), max-iter: 1000/1000, Adam, lr: 0.001, batch: 4, loss log
1DCNN	Large	Adam, lr: 0.001, loss: cross-entropy, dropout:0.1, batch-size: 16, epochs: 100
	Base	Adam, lr: 0.001, loss: cross-entropy, dropout: 0.1, batch-size: 16, epochs: 100

2.4 Evaluation Metrics

The evaluation of the proposed method was conducted using standard classification metrics, including accuracy, sensitivity, specificity, and F1 score. Accuracy measures the model's overall ability to correctly classify speech samples into both classes (positive and negative). Sensitivity, also known as the true positive rate (TPR), is the probability that the model correctly identifies positive samples, given that the individual is truly positive. Specificity, or the true negative rate (TNR), is the probability that the model correctly identifies negative samples, given that the individual is truly negative. The F1 score is the harmonic mean of precision and recall, providing a balanced metric that accounts for both false positives and false negatives, particularly useful in cases of class imbalance.

3 Results

Results reported in this section are from an independent test set; for all the experiments, a manual seed was applied to ensure the same test set was used for fair performance comparisons between the models. Table 2 reports performance metrics for different ML algorithms and our proposed model using the Base and Large Whisper encoder. Rows represent different ML algorithms, and the columns represent metric results using both Base and Large encoders. Bold values represent the highest performance between ML algorithms, while bold and underlined denote the best performance achieved. Overall, all ML algorithms demonstrated good performance. The accuracy metrics range from 87.64% to 95.51%, indicating the Whisper encoder's capabilities to capture dysphonic-related features in speech samples.

Table 2. Test set results for different ML algorithms.

ML	Accuracy		F1 score		Sensitivity		Specificity	
	Base	*Large*	*Base*	*Large*	*Base*	*Large*	*Base*	*Large*
SVM-rbf	88.76	93.26	0.912	0.947	88.14	91.53	90.00	96.67
SVM-linear	88.76	93.26	0.912	0.947	88.14	91.53	90.00	96.67
RF	87.64	92.13	0.901	0.938	84.75	89.83	93.33	96.67
MLP	89.89	91.01	0.922	0.935	89.83	98.31	90.00	76.67
1DCNN	**92.13**	**95.51**	**0.937**	**0.966**	**88.14**	**94.92**	**100**	**96.67**

As seen in the table, in the Base model, all algorithms except 1DCNN achieved comparable performance, with the lowest accuracy of 87.76% by RF and the highest accuracy of 89.89% by MLP. Notably, both SVM-rbf and SVM-linear achieved the same performance across all metrics, indicating that the choice of kernel type does not significantly impact the results.

We observed performance improvements across all ML algorithms when comparing the results achieved using the Large Whisper model to the Base model. The Large model consistently delivered better accuracy and F1 scores in all ML algorithms. The difference in performance between the two models ranges from 1.12% to 4.5% in terms of accuracy in MLP and SVM algorithms, respectively. The proposed 1DCNN model trained on speech embeddings from a Large Whisper encoder achieves the highest accuracy, illustrating the impact of the Larger model characterized by a greater number of transformer layers and trainable parameters. This scalability improves the models' capacity to capture more distinctive features in dysphonia-related speech samples.

When comparing the results of our proposed 1DCNN architecture to off-the-shelf ML algorithms, we observe that our architecture consistently outperforms other algorithms in both variations despite the limited training data available. The model achieves an accuracy of 95.51% and an F1 score of 0.966, demonstrating outstanding performance in dysphonia detection.

Confusion Matrices A and B from Fig. 3 illustrate the performance of the proposed method with Base and Large Whisper encoders in classifying the speech sample into two classes: HC (Healthy Control) and Dys (Dysphonia), respectively. Columns represent the actual labels, while rows represent the predicted labels by the model. The values in the matrix starting from the top left represent True Negative (TP), False Positive (FP), False Negative (FN), and True Negative (TN), respectively. The diagonal values indicate the number of correct predictions. Summing the values in each column will provide the total number of samples in that particular class in the test set.

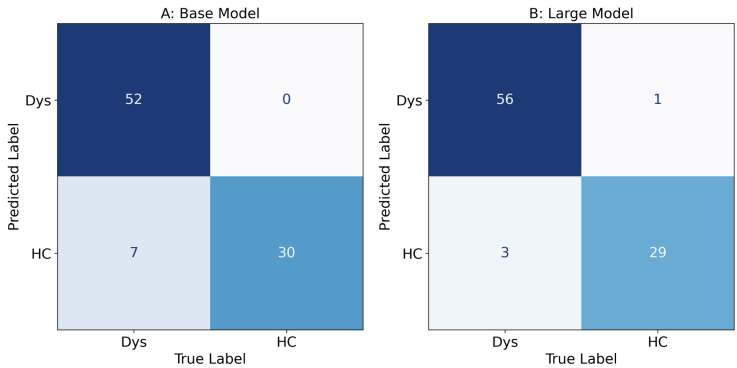


Fig. 3. Confusion matrix of 1DCNN model with Base and Large Whisper encoder.

From the analysis of the confusion matrices, it is evident that speech embedding from the Whisper Large encoder approach outperforms Base. It shows higher accuracy in correctly predicting instances. FN indicates that the model predicted the speech sample as healthy, but in reality, it is dysphonic speech. In this case, the Base model has 7 FN, whereas the Large model has only 3 FN.

This indicates that the Larger encoder is more effective at correctly identifying patterns for distinguishing between two classes, highlighting its superiority in capturing speech characteristics associated with dysphonia.

Moreover, we also conducted multiclass classification. In this scenario, we trained ML algorithms using extracted speech embeddings to distinguish between HC, OD, and FD. When evaluating the results in Table 3, it is clear that the performance of the Large and Base speech encoder models aligns closely across all evaluation metrics. Only SVM-rbf and RF with the Large Whisper encoder achieve slightly better performance compared to their Base counterparts. In contrast, the MLP performs worse with the Large model. These results suggest that both models capture similar speech characteristics for distinguishing dysphonia categories. Consequently, it can be inferred that the complexity of the larger model does not grant a significant advantage over the Base model for this specific classification task.

Table 3. Multiclass Test set results for different ML algorithms.

ML	Accuracy		F1 score		Sensitivity		Specificity	
	Base	*Large*	*Base*	*Large*	*Base*	*Large*	*Base*	*Large*
SVM-rbf	71.91	73.03	0.690	0.667	69.80	69.76	85.65	85.87
SVM-linear	70.79	70.79	0.673	0.649	68.47	67.80	85.04	84.83
RF	69.66	73.03	0.612	0.685	65.99	70.47	84.13	85.99
MLP	69.66	67.42	0.648	0.661	66.92	66.10	84.30	83.33
1DCNN	**76.40**	**76.40**	**0.751**	**0.753**	75.14	**75.62**	**88.00**	**88.12**

The confusion matrices from Fig. 4 compare the classification performance using the Base and Large Whisper models for HC, FD, and OD. The Large model demonstrates a slight improvement in correctly classifying HC (28 vs. 27). However, it performs slightly worse in correctly classifying FD samples (25 vs. 27) and exhibits a higher misclassification rate for FD, with more FD samples being predicted as OD. Both models struggle to distinguish between OD and FD. They predict 8 samples as FD while they are OD, likely due to overlapping acoustic features. Overall, the Large model provides minor improvements, particularly in HC and OD classification, while performing worse in classing FD.

Comparing the results obtained from our proposed method in binary and multiclass classification tasks with other machine learning classifiers, we observe that the proposed 1DCNN significantly outperforms other algorithms, achieving an almost 9% absolute difference compared to MLP when using speech encoded from the Larger Whisper. These findings highlight the superior ability of 1DCNN to effectively handle complex, encoded speech data. Notably, 1DCNN maintains consistent performance with both Base and Large models, unlike MLP, which demonstrates a decrease in accuracy with the Larger model.

Comparing these results with other studies, we examine the outcomes of three recent publications that employed the same datasets. Figure 5 illustrates

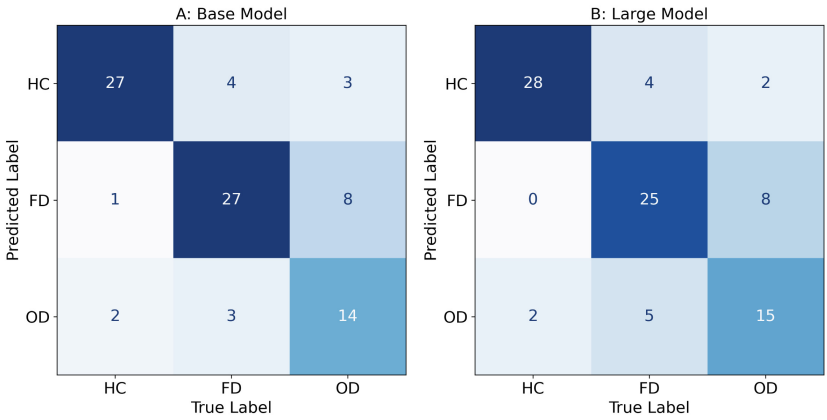


Fig. 4. Confusion matrix of multiclass classification.

comparisons of our best-performing model with related works. The x-axis represents different methods employed in the same dataset, and the y-axis represents achieved performance.

Our proposed method clearly outperforms previous works by a large margin, especially in the case of the Large encoder. Our 1DCNN with Large Whisper speech encoder achieves more than 95% accuracy and F1 score, which is an absolute improvement of more than 4%, 5%, and 9% compared with the results achieved by [4,5,28], respectively.

Moreover, in the multiclass classification comparison with other previous work, only [4] performed three-class classification. Looking at the Table 3, in the case of using the Base Whisper, only 1DCNN outperforms previous related work. Although the Larger Whisper model leads to improved performance in SVM-Linear and RF, they do not outperform the previous approach. Figure 6 illustrates a comparison of our best-performing algorithms with previous approaches. As can be seen from Fig. 6, our proposed method outperforms previous related work, achieving improvements of more than 3% and 6% in both accuracy and F1 score, respectively.

To summarize our findings, the proposed 1DCNN model utilizing the Large and Base Whisper speech encoder significantly outperforms previous methodologies in both binary and multiclass dysphonia classification tasks. The results demonstrate the Whisper encoder's efficacy in capturing dysphonic features, with the Large model consistently yielding higher performance in all classification metrics across all ML algorithms. These findings highlight the effectiveness of Whisper's speech embeddings combined with 1DCNN in capturing dysphonia-related characteristics and improving dysphonia detection performance.

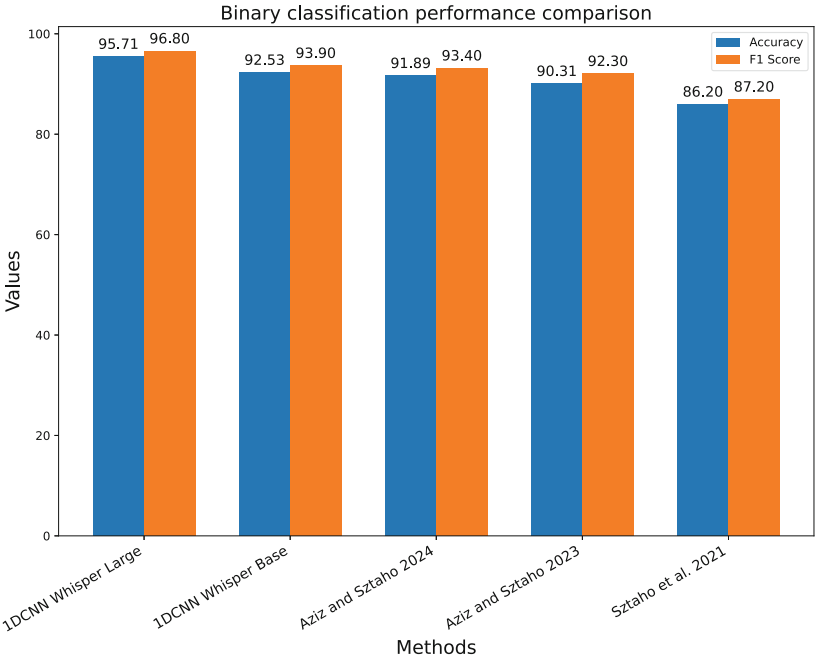


Fig. 5. Performance comparison of the proposed method with other related works in binary classification.

4 Discussion

Transfer learning has been shown to have a significant impact on improving accuracy in domains where the available dataset is sparse. Several approaches have proposed the use of transfer learning for diagnosing dysphonia, but they have primarily focused on computer vision models, which may not perform optimally due to the mismatch between visual data and speech tasks. In this paper, we adopted the transfer learning approach using the Whisper encoder, a robust speech processing architecture trained on extensive speech datasets, to capture speech characteristics at the utterance level. Our results demonstrated the capabilities of the Whisper encoder for diagnosing dysphonia.

Our results from binary classification show that the Large variant of the Whisper model often leads to better accuracy across all ML algorithms. However, utilizing the Large variant does not increase classification performance when classifying HC, OD, and FD. This indicates that a larger number of transformer layers and model parameters capture similar features related to distinguishing between dysphonia subtypes and do not provide a more discriminative feature set. Another reason might related to the limited number of speech samples in the training dataset, which is not sufficient to capture more diverse features for better distinguishing between FD and OD. Our proposed 1DCNN architecture outperforms other ML algorithms utilizing both Base and Large Whisper models

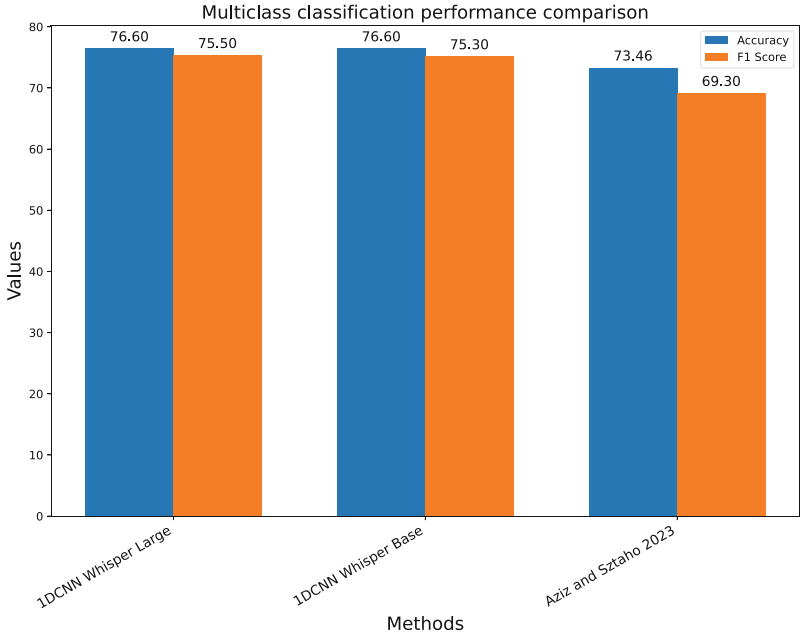


Fig. 6. Performance comparison of the proposed method with other related works in multiclass classification.

in binary and multiclass experiments, highlighting the superior capabilities of 1DCNN in processing speech embeddings compared to other ML methods.

Comparing the results achieved with previous related works, our proposed method employing both Base and Large Whisper models significantly outperforms other related work in binary and multiclass classification. This highlights the capabilities of Whisper models combined with 1DCNN to distinguish between healthy and dysphonic speech as well as between different dysphonia categories.

The improved performance of our proposed method, especially in binary classification, demonstrates significant potential for clinical applications. Enhanced accuracy in identifying dysphonia can offer healthcare providers more precise diagnostic tools for pre-screening patients, potentially leading to earlier interventions and personalized treatment plans, thus improving patient outcomes.

This approach can be integrated into early diagnostic phases, such as during general practitioner visits or through mobile devices at home, offering a cost-effective complement to clinical expertise. While clinicians may benefit from these computational techniques, their role should be to augment, not replace, professional judgment in diagnosis.

Although the proposed approach achieves promising results, it also has some limitations. The dataset used in this study is relatively small and may not capture

all the diversity in the dysphonia population. Future studies should validate the model with larger and more diverse datasets to ensure its robustness and applicability across different populations.

5 Conclusion

In this paper, we adopted the Whisper speech encoder for binary and multiclass classification of dysphonic speech and proposed a custom 1DCNN. Our findings from the binary classification task show that utilizing the Larger Whisper encoder variant performed better than the Base counterpart, outperforming previous approaches on the same dataset. While the Large model outperformed the Base in binary classification, our results from multiclass classification imply that both models of Whisper exhibit similar performance. This suggests that the additional transformer layers in the Large model do not capture more discriminative features related to dysphonia categories. Moreover, our custom 1DCNN model outperformed classical ML algorithms, even with a limited number of speech samples. The proposed method achieved 95.51% accuracy in binary classification and 76.40% accuracy in multiclass classification, significantly surpassing previous approaches using the same dataset. This highlights the superior capability of our 1DCNN model in distinguishing between healthy and dysphonic speech, as well as differentiating between dysphonia categories.

In conclusion, the Whisper speech encoder, particularly when combined with our custom 1DCNN architecture, shows significant promise in improving the accuracy of dysphonia classification. Future research could explore the optimization of the model architecture further and adopt it in multilingual scenarios by training the model width datasets from diverse languages.

Acknowledgments. This work was partly funded by project no. K143075, which has been implemented with the support provided by the National Research, Development, and Innovation Fund of Hungary, financed under the K_22 funding scheme.

References

1. Amami, R., Amami, R., Trabelsi, C., Mabrouk, S.H., Khalil, H.A.: A Robust Voice Pathology Detection System Based on the Combined BiLSTM–CNN Architecture. 1. **29**(2), 202–210 (2023). https://doi.org/10.13164/mendel.2023.2.202
2. Angelillo, I.F., Di Maio, G., Costa, G., Angelillo, I.F., Barillari U.: Prevalence of occupational voice disorders in teachers. J. Prev. Med. Hyg. **50**(1), (2009). https://doi.org/10.15167/2421-4248/jpmh2009.50.1.152
3. Aronson, A.E.: Clinical Voice Disorders: An Interdisciplinary Approach. Thieme (1990)
4. Aziz, D., David, S.: Multitask and transfer learning approach for joint classification and severity estimation of dysphonia. IEEE J. Transl. Eng. Health Med. **12**, 233–244 (2023). https://doi.org/10.1109/JTEHM.2023.3340345

5. Aziz, D., Sztahó, D.: Dysphonia detection using a fully convolutional neural network adapted to dynamic speech lengths. In: 2nd Workshop on Intelligent Infocommunication Networks, Systems and Services (WI2NS2) (2024). https://doi.org/10.3311/WINS2024-003
6. Chen, Z., Zhu, P., Qiu, W., Guo, J., Li, Y.: Deep learning in automatic detection of dysphonia: Comparing acoustic features and developing a generalizable framework. Int. J. Lang. Commun. Disord. **58**(2), 279–294 (2023). https://doi.org/10.1111/1460-6984.12783
7. Crevier-Buchman, L., Ch, T., Sauvignet, A., Brihaye-Arpin, S., Monfrais-Pfauwadel, M.C.: Diagnosis of non-organic dysphonia in adult. Revue de Laryngologie-Otologie-Rhinologie **126**(5), 353–360 (2005)
8. Dankovičová, Z., Sovák, D., Drotár, P., Vokorokos, L.: Machine learning approach to dysphonia detection. Appl. Sci. **8**(10), 1927 (2018). https://doi.org/10.3390/app8101927
9. El Emary, I.M.M., Fezari, M., Amara, F.: Towards developing a voice pathologies detection system. J. Commun. Technol. Electron. **59**(11), 1280–1288 (2014). https://doi.org/10.1134/S1064226914110059
10. Fagherazzi, G., Fischer, A., Ismael, M., Despotovic, V.: Voice for health: the use of vocal biomarkers from research to clinical practice. Digital Biomarkers **5**(1), 78–88 (2021). https://doi.org/10.1159/000515346
11. Favaro, A., et al.: Interpretable speech features vs. DNN embeddings: What to use in the automatic assessment of Parkinson's disease in multi-lingual scenarios. Comput. Biol. Med. **166**, 107559 (2023). https://doi.org/10.1016/j.compbiomed.2023.107559
12. Harar, P., Alonso-Hernandezy, J.B., Mekyska, J., Galaz, Z., Burget, R., Smekal, Z.: Voice Pathology Detection Using Deep Learning: a Preliminary Study. In: 2017 International Conference and Workshop on Bioinspired Intelligence (IWOBI), pp. 10–12. IEEE (2017). https://doi.org/10.1109/IWOBI.2017.7985525
13. Harar, P., Galaz, Z., Alonso-Hernandez, J.B., Mekyska, J., Burget, R., Smekal, Z.: Towards robust voice pathology detection. Neural Comput. &. Applic. **32**(20), 15747–15757 (2020). https://doi.org/10.1007/s00521-018-3464-7
14. Islam, R., Tarique, M.: A novel convolutional neural network based dysphonic voice detection algorithm using chromagram. Inter. J. Elect. Comput. En. (2088-8708) **12**(5) (2022). https://doi.org/10.11591/ijece.v12i5.pp5511-5518
15. Johns, M.M., Sataloff, R.T., Merati, A.L., Rosen, C.A.: Article commentary: Shortfalls of the american academy of otolaryngology–head and neck surgery's clinical practice guideline: Hoarseness (dysphonia). Otolaryngology-Head and Neck Surgery **143**(2), 175–177 (2010). https://doi.org/10.1016/j.otohns.2010.05.026
16. de Jong, F.I.C.R.S., Kooijman, P.G.C., Thomas, G., Huinck, W.J., Graamans, K., Schutte, H.K.: Epidemiology of Voice Problems in Dutch Teachers. Folia Phoniatr. Logop. **58**(3), 186–198 (Apr 2006). https://doi.org/10.1159/000091732
17. Jothilakshmi, S.: Automatic system to detect the type of voice pathology. Appl. Soft Comput. **21**, 244–249 (2014). https://doi.org/10.1016/j.asoc.2014.03.036
18. Martínez, D., Lleida, E., Ortega, A., Miguel, A., Villalba, J.: Voice pathology detection on the saarbrücken voice database with calibration and fusion of scores using multifocal toolkit. In: Torre Toledano, D., et al. (eds.) IberSPEECH 2012. CCIS, vol. 328, pp. 99–109. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-35292-8_11
19. Nerrière, E., Vercambre, M.N., Gilbert, F., Kovess-Masféty, V.: Voice disorders and mental health in teachers: a cross-sectional nationwide study. BMC Public Health **9**(1), 1–8 (2009). https://doi.org/10.1186/1471-2458-9-370

20. Park, H.J., Shin, B.J.: Usefulness of glottal inverse filtering analysis in pathological voice1. J. Speech **30**(1), 041–048 (2021)
21. Pedregosa, F., et al.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011). https://doi.org/10.48550/arXiv.1201.0490
22. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust Speech Recognition via Large-Scale Weak Supervision. arXiv (Dec 2022). https://doi.org/10.48550/arXiv.2212.04356
23. Ravanelli, M., et al.: SpeechBrain: A general-purpose speech toolkit arXiv:2106.04624 (2021)
24. Reid, J., Parmar, P., Lund, T., Aalto, D.K., Jeffery, C.C.: Development of a machine-learning based voice disorder screening tool. Am. J. Otolaryngol. **43**(2), 103327 (2022). https://doi.org/10.1016/j.amjoto.2021.103327
25. Ribas, D., Pastor, M.A., Miguel, A., Martínez, D., Ortega, A., Lleida, E.: Automatic voice disorder detection using self-supervised representations. IEEE Access **11**, 14915–14927 (2023). https://doi.org/10.1109/ACCESS.2023.3243986
26. Roy, N., Merrill, R.M., Thibeault, S., Parsa, R.A., Gray, S.D., Smith, E.M.: Prevalence of Voice Disorders in Teachers and the General Population. ASHA Wire (Apr 2004). https://pubs.asha.org/doi/10.1044/1092-4388%282004/023%29
27. Stachler, R.J., et al.: Clinical practice guideline: Hoarseness (dysphonia) (update). Otolaryngology–Head and Neck Surgery **158**(S1), S1–S42 (2018).https://doi.org/10.1177/0194599817751030
28. Sztahó, D., Kiss, G., Tulics, M.G.: Deep learning solution for pathological voice detection using lstm-based autoencoder hybrid with multi-task learning. In: BIOSIGNALS, pp. 135–141 (2021). https://doi.org/10.5220/0010193101350141
29. Tulics, M.G., Vicsi, K.: The automatic assessment of the severity of dysphonia. Int. J. Speech Technol. **22**(2), 341–350 (2019). https://doi.org/10.1007/s10772-019-09592-y
30. unknown: Voice disorders. https://www.hopkinsmedicine.org/health/conditions-and-diseases/voice-disorders
31. Vaswani, A., et al.: Attention Is All You Need. arXiv (Jun 2017). https://doi.org/10.48550/arXiv.1706.03762
32. Verde, L., De Pietro, G., Sannino, G.: Voice disorder identification by using machine learning techniques. IEEE Access **6**, 16246–16255 (2018). https://doi.org/10.1109/ACCESS.2018.2816338
33. Williams, N.R.: Occupational groups at risk of voice disorders: a review of the literature. Occup. Med. **53**(7), 456–460 (2024). https://doi.org/10.1093/occmed/kqg113
34. van der Woerd, B., et al.: A machine-learning algorithm for the automated perceptual evaluation of dysphonia severity. J. Voice (2023). https://doi.org/10.1016/j.jvoice.2023.06.006

Approach to Assessing the Quality of Syllable Pronunciation by Patients in the Process of Speech Rehabilitation Based on Comparison with Healthy Speakers

German Egle, Dariya Novokhrestova(✉), Svetlana Tomilina, and Evgeny Kostyuchenko

Tomsk State University of Control Systems and Radioelectronics, Lenina Street 40, 634050 Tomsk, Russia
ndi@fb.tusur.ru
http://www.tusur.ru

Abstract. The article proposes a modification of the methodology for assessing the quality of pronunciation of a patient's syllables in the process of speech rehabilitation. A modification of the technique consists in applying an approach that allows the use of recordings of healthy speakers to assess the quality of pronunciation of various phonemes. To assess the quality of pronunciation, we use the distance between two files with recordings in wav format; the smaller the distance between two files, the more similar they are considered. Distance calculation methods discussed included DTW, EDR, ERP, LCSS, MSM, and Euclidean distance. The accuracy of the method was tested using values obtained using the instantaneous energy envelope and the Gilbert envelope, in the first case ERP was the most accurate method, in the second EDR. It was revealed that the LCSS and ERP methods perform calculations much longer than other methods and do not have high accuracy, as a result, they are not included in the list of the best calculation methods. The hypothesis that the two means were equal was also tested using the Mann-Whitney method to confirm that the distance between the healthy speaker recording and the preoperative patient recording differs from the distance between the healthy speaker recording and the recording of the patient after surgery. As a result, it turned out that the most accurate and fastest method for calculating the distance between two records is EDR. Analysis of the results obtained showed that this approach is applicable to solving the problem of speech analysis, but requires significant improvements and subsequent research due to the low accuracy of the put forward theories.

Keywords: Speech Assessment · Speech Rehabilitation · Healthy Speaker · Assessment Algorithm

A. Karpov and V. Delić (Eds.): SPECOM 2024, LNAI 15299, pp. 367–376, 2025.
https://doi.org/10.1007/978-3-031-77961-9_27

1 Introduction

Statistics of oncological diseases of the organs of the speech tract [1] show that the number of cases with this localization does not decrease every year. Therefore, research and development of speech rehabilitation methods and speech assessment algorithms in medicine remain relevant areas now. Assessment of the speech quality in rehabilitation allows you to track the degree of speech recovery after surgery and quantify speech intelligibility, and as a result, adjust the treatment method of an individual patient [2] if it needs.

In accordance with the clinical recommendations adopted at the moment, the assessment of pronunciation and speech quality in the process of speech rehabilitation does not have a single methodology, and the available instructions are based on the speech therapist expert assessment, namely assessment by listening to the patient speech. Previously, a speech assessment technique based on an automated algorithmic comparison of syllable pronunciation records of the same patient at different stages of speech rehabilitation was proposed by the authors and implemented in some medical institutions. As a reference (the state of speech to which it is necessary to return in the process of rehabilitation), the patient recordings of speech before surgical intervention in the organs of speech production are used. The available technique makes it possible to evaluate different recordings based on the calculation of the similarity metric between the patient's recordings before surgery with recordings after rehabilitation. This provides a quantitative assessment of speech restoration during the patient's speech rehabilitation process [3]. In this methodology, speech assessment means the calculation of two indicators: an assessment of the pronunciation of syllables to assess the correctness of the pronunciation of individual minimum units of speech and an assessment of the pronunciation of phrases to assess the permeability of speech. This study will focus specifically on the assessment of syllable pronunciation.

However, this technique has a significant drawback: the need for reference records for each patient, which will be considered as high-quality speech. High-quality speech means speech with correct pronunciation of the phonemes and other speech units of the Russian language, taking into account the original features of the speaker's speech. That is, if a person initially has rhotacism, his speech is considered high-quality for the purpose of speech rehabilitation assessment of his case. Because within the framework of speech rehabilitation, there is no goal to achieve ideal pronunciation of phonemes, but the task is to return to the preoperative speech level. The specificity of the studied disease implies that in some cases, at the time of admission of the patient to the hospital, his speech is already not high-quality, distorted by the presence of neoplasms. And for such patients there is no way to record high-quality speech, which will become a reference for evaluation in the rehabilitation process. Therefore, it is proposed to refine the methodology within the framework of an approach to assessing the patient's speech based on a comparison with the speech of a group of healthy speakers. This work is an attempt to exactly expand the methodology for assessing the pronunciation of syllables, therefore the algorithms used (signal processing, calculation of similarity metrics) and the recorded and analyzed speech material (list of recorded syllables and list of problematic phonemes) are the initial parameters and are not studied in this work.

2 Proposed Approach

The assessment according to the proposed approach is calculated as the average value of the distances between the recording of a syllable pronounced by the patient and recordings of the same syllable pronounced by healthy speakers. The proposed method takes into account the fact that a person can pronounce the same unit of speech in different ways and it is to account for this error that not one healthy speaker is used, but several healthy speakers.

According to the calculation algorithm, an audio recording containing the pronunciation of one syllable is converted into a sequence of numeric values. To test the applicability approach, we indicate that the patient's records may belong to three "intermediate points" in the treatment and rehabilitation process: "before surgery" - speech was recorded before surgery, "after surgery" - speech was recorded after surgical intervention, but before rehabilitation procedures, "after rehabilitation" - speech was recorded after a full or partial set of speech restoration training. The algorithm involves calculating the metric between two sequences. The metric is selected based on two criteria: accuracy and performance (calculation speed). The accuracy of the method must be determined based on their hypothesis that the distance from the recording of a healthy speaker to the recording of a patient "before surgery" it will be less than the distance from the recording of a healthy speaker to the recording of the patient "after surgery". According to the original methodology, it is assumed that speech before surgery is of the highest quality (intelligible, if we use the definitions of speech therapists), speech after surgery is of the lowest quality, and during the rehabilitation process speech should become of higher quality and approach the preoperative level (but in most cases not reach it). This approach to assessing the level of speech quality is confirmed by oncologists and speech therapists based on observation data of patients.

The method that shows the most correct results is considered the most accurate. The calculation speed is defined as the time required to find the distances between all the syllable entries pronounced by the patient and all the healthy speaker entries. Accuracy is a more important criterion, as it shows the applicability of this metric in the approach, however, the speed of calculations is also being investigated in view of the need for an assessment within a limited period of time (as part of a patient's session with a speech therapist or oncologist).

2.1 Dataset of Speech Recordings

Two sets of recordings were generated to assess the applicability of the approach: a set of patient recordings and a set of healthy speaker recordings. Each recording is an audio file in mono format with.wav extension with a sampling rate of 12,000 Hz with a recording of the pronunciation of one syllable.

Each patient has at least one recording session in each of the rehabilitation stages. A session means a set of recordings the pronunciation of syllables according to the list of syllables for recordings, which consists of 30 syllables: 10 syllables containing each of the problematic phonemes at the beginning of the syllable. This set of syllables obviously does not cover all linguistic units of the Russian language, but it takes into account the most problematic phonemes for the considered localization of the disease

of the patients studied. A problematic phoneme is the phoneme that most often changes in pronunciation after surgical intervention in the organs of speech production. Each of the problematic phonemes occurs in a given set of syllables at least 15 times. Patient recordings were made by speech therapist or oncologist at Blokhin Research Institute (Moscow, Russia). All patients were diagnosed with tongue cancer and one of two surgical options: hemiglossectomy without tongue reconstruction and hemiglossectomy with tongue reconstruction. The study set of patient records contains records of 27 patients.

The set of recordings of healthy speakers also consists of recordings of syllable pronunciation sessions according to the same list (for the possibility of comparing the same syllables in the patient and healthy speakers). The database of recordings of healthy speakers consists of 14 people: 7 women and 7 men, the age of the speakers is in the range from 35 to 65 years. These recordings were selected from an existing set of recordings of syllables from healthy speakers in accordance with the age range of the patients (the age range of the speakers coincides with the age range of the patients). This range also coincides with the most common age of patients with oncological diseases of the speech apparatus is just the age from 35 to 65 years [4]. In this study, only existing recordings was used. If the possibility of using this approach to assessment is confirmed, the need to expand the set of audio recordings of healthy speakers of the required age will be considered and the issue of forming a criterion for selecting a list of healthy speakers for various situations will be investigated.

2.2 Algorithm for Evaluating the Speech Quality

To assess the quality of the patient's speech, an algorithm was developed as shown in Fig. 1. This algorithm was developed taking into account its possible inclusion in the speech assessment methodology.

One patient session (30 records) and all records of the database of healthy speakers are submitted for input. For each recording with a syllable, the distances from the corresponding file with each of the recordings of the same syllable of all healthy speakers are calculated. After it, the average value of distances for each syllable is calculated. This set of values is assessments for different syllable in session. A set of assessments is entered into the database, and the average of these values is calculated as the session score. The assessment of the session numerically reflects the quality of the patient's speech as part of the assessment of the pronunciation of syllables.

2.3 Similarity Metrics and Distance Measures

To implement the approach into the methodology, it is necessary to determine the best method (metric or measure) for finding the distance between two records based on comparing the values obtained using metrics that were studied earlier as part of the construction of the initial methodology. The following methods of finding the distance between two sequences were chosen: DTW, ERP, EDR, MSM, LCSS and Euclidean distance. 4 of the 5 proposed metrics are distance metrics, which are interpreted as "the smaller the value, the more similar the sequences", the LCSS metric is a similarity metric (the more values, the more similar the sequences).

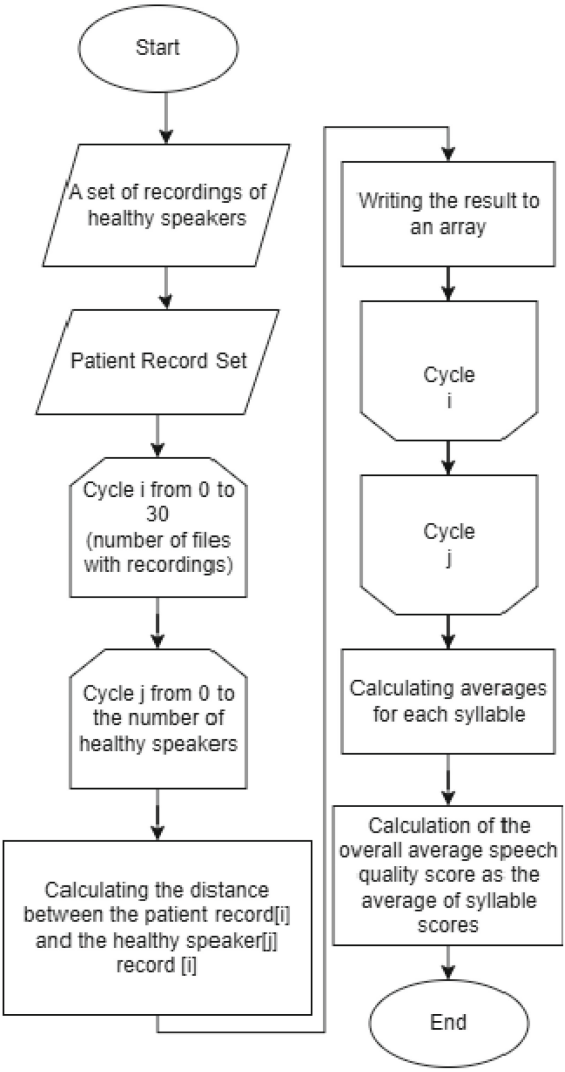


Fig. 1. Algorithm for evaluating the quality of speech.

To calculate the values of the estimates, programs were written in Python [5], each program calculates the distance between 14 healthy speakers and all records of 27 patients using DTW, ERP, EDR, MSM, LCSS, and Euclidean distance methods.

- DTW (Dynamic Time Warping) is an algorithm for dynamic transformation of the timeline: This method is used to compare time frames of different lengths, allowing you to calculate the distance between them, taking into account pos-sible distortions in the timeline [6].

- ERP (Edit Distance with Real Penalty) - An algorithm for editing the distance using a penalty: This method calculates the minimum number of operations (insert, delete, replace) required to transform one row into another, used to compare sequences [7].
- EDR (Edit Distance on Real Sequences) - Edit distances for real sequences: This method is similar to the ERP algorithm, but it is designed to work with re-al sequences of numbers, not just text [8].
- MSR (Multi-Scale Matching) - Multiscale matching: This method allows you to compare the two time series, taking into account different time scales and amplitude changes [9].
- LCSS (Longest Common Sequence) is the longest common subsequence: This method finds the largest common sequence of elements between two sequences [10].
- The Euclidean distance is a classical metric that measures the Euclidean distance between two points in n–dimensional space and is often used to compare between points in the function space [11].

All calculations carried out by the programs used 2 different envelopes of the audio signal: the instantaneous energy envelope and the Gilbert envelope.

- Instantaneous Energy Envelope [12] is a signal analysis method that is used to study changes in signal energy over time. This method involves calculating the instantaneous power of the signal, then integrating the instantaneous power to produce instantaneous energy, and finally isolating the envelope of the received instantaneous energy signal. This method is often used in acoustics, speech processing, and equipment diagnostics.
- The Gilbert Envelope [13] is a signal analysis method that is used to extract information about the amplitude and phase of a signal. The Gilbert envelope is the envelope of a signal constructed using the Gilbert frequency transform, which is a mathematical transformation used to represent a signal as a complex number. Then, the module of the complex number is extracted to obtain information about the amplitude of the signal, and the argument of the complex number to obtain information about the phase of the signal. First, the signal undergoes a Gilbert transformation, and then the envelope is obtained by combining the original signal and the resulting transformation. This method is often used in telecommunications, signal processing and automatic control.

Currently, both types of envelopes are used to calculate quantitative estimates using a method for assessing the quality of syllable pronunciation, without a clear preference for one of them. Therefore, it is not possible to select only one type of envelope for analysis in this study.

3 Results

3.1 The Metrics Results

To determine the most accurate metric in algorithm for calculating the distance between two recordings, 2 hypotheses are tested. The first hypothesis looks like this: the distance between the recording of a healthy speaker and the recording of the patient before surgery

is less than the distance between the recording of a healthy speaker and the recording of the patient after rehabilitation, and the largest value is the distance between the recording of a healthy speaker and the recording of the patient after surgery. The second hypothesis is a simplification of the first assumption: the distance between the recording of a healthy speaker and the patient before surgery is less than the distance between the recording of a healthy speaker and the recording of the patient after surgery. The second hypothesis was introduced for analysis after studying audio recordings by speech specialists, who determined that it was impossible to clearly establish the fact of speech improvement after rehabilitation measures due to the characteristics of the disease.

As a result of the analysis of the obtained values, the most accurate methods for using the instantaneous energy envelope were identified. Results were analyzed for each of hypothesis, and also with dividing all patients into two groups by gender and by surgical intervention. Results for algorithm with instantaneous energy envelope are shown in Table 1. Results for algorithm with Gilbert envelope are shown in Table 2. The tables present only the best (most accurate) calculated metrics, as well as the Percentage of agreement between quantitative values and the proposed hypothesis for the best metric. Surgical intervention 1 is hemiglossectomy without tongue reconstruction, and Surgical intervention 2 is hemiglossectomy with tongue reconstruction.

Table 1. Results of checking the accuracy for algorithm with an envelope of instantaneous energy.

All data		Divided by gender				Divided by diagnosis			
Hypothesis 1	Hypothesis 2	Hypothesis 1		Hypothesis 2		Hypothesis 1		Hypothesis 2	
		Male	Female	Male	Female	Surgical intervention 1	Surgical intervention 2	Surgical intervention 1	Surgical intervention 2
ERP	ERP	ERP	ERP	EDR	ERP	ERP	LCSS	ERP	LCSS
17%	50%	15%	22%	55%	65%	30%	15%	20%	22%

Table 2. Results of checking the accuracy for algorithm with the Gilbert envelope.

All data		Divided by gender				Divided by diagnosis			
Hypothesis 1	Hypothesis 2	Hypothesis 1		Hypothesis 2		Hypothesis 1		Hypothesis 2	
		Male	Female	Male	Female	Surgical intervention 1	Surgical intervention 2	Surgical intervention 1	Surgical intervention 2
EDR	EDR	EDR	EDR	EDR	DTW	EDR	EDR	Euclid	EDR
21%	59%	19%	27%	64%	70%	27%	19%	27%	19%

Based on the data presented in the table, it can be concluded that the MSM method has never shown the best result in accuracy, as a result in the future It is not considered in research. The LCSS and ERP method performed distance calculations between all files of healthy speakers and all files with patient records of 14 patients in about 5–6 h, the remaining methods performed calculations in an average of 10 min, as a result, LCSS

and ERP distance calculation methods are not considered in further research, since their calculation speed indicator is low.

Since the study showed that the ERP method is not suitable in terms of speed, the Gilbert envelope is further used, where the methods that have shown high accuracy are EDR, DTW and Euclidean separation. It is obvious that even the "best" metrics show a rather low agreement accuracy, which shows the need to refine this approach to its further use in the methodology for quantitative assessment of speech quality.

3.2 Hypothesis About the Median Difference

A proposal was put forward to test the applicability of this approach based on an analysis of the statistical significance of the differences in the sample of distance values in pairs of various combinations of "healthy speaker" and "patient at various stages of rehabilitation.". To test the hypothesis of the median difference between the sample of healthy speakers and the samples between healthy speakers and patients before surgery, healthy speakers and patients after surgery, and healthy speakers and patients after rehabilitation, the Mann-Whitney statistical criterion was tested. The sample between healthy speakers was formed as follows: the distance between the first and second speaker, between the first and third speaker, and so on was calculated. As a result, distance data was obtained between 30 recordings of fourteen speakers using DTW, EDR, MSM, Euclidean distance methods, in total, the sample consisted of 5,278 values using one distance calculation method. SPSS Statistics, a computer program for statistical data processing, was used to calculate the Mann-Whitney criterion. To calculate the criterion, a sample of healthy speakers consisting of 5278 values was submitted to the program input, as well as a sample of values between healthy speakers and the patient in three variants - before surgery, after surgery and after rehabilitation. In accordance with the Mann-Whitney test, the null hypothesis of equality of medians was formulated. As a result, 14 results were obtained, each of which showed that all combinations of samples at a significance level of 0.05 were different, since the null hypothesis was rejected. As a result, the medians of all samples are different. However, the analysis also revealed that when considering the range of distance values between healthy speakers and the range of distance values between a healthy speaker and a patient before surgery, these ranges are intersecting. That is, it cannot be stated that the distance between healthy speakers is always less than the distance between a healthy speaker and a patient before surgery. From all these results we can conclude that this calculation technique makes it possible to obtain different quantitative estimates of speech for each state of speech of speakers and patients.

4 Conclusion

As a result of the study, the approach for modification of method for assessing the quality of speech was proposed. This approach is about an assessment of the quality of speech by finding the distances between the patient's syllable recordings and the recordings of syllable of healthy speakers. A database of recordings of healthy speakers was collected, consisting recordings of 14 people: 7 women and 7 men aged 35 to 65 years. The following calculation metrics and measures were chosen to calculate the

distances between the patient's records and the recordings of healthy speakers and: DTW, EDR, ERP, MSM, LCSS, Euclidean distance. Two types of envelope were chosen for analysis: the instantaneous energy envelope and the Gilbert envelope.

As a result of the analysis, summary tables were compiled reflecting the accuracy of each of the selected methods for calculating the distances between two records. It was found that when using the instantaneous energy envelope, ERP is the most accurate method, and when using the Gilbert envelope, EDR is the most accurate method. The distance calculation time was also taken into account, as a result, the LCSS methods and the ERP method showed a calculated time of results 35 times longer, compared with other distance calculation methods, as a result, the ERP and LCSS methods were not considered in the further study.

A hypothesis was put forward about the median difference samples values in pair "healthy speaker–healthy speaker" and in pair "healthy speaker–patient". To confirm the hypothesis put forward, the Mann-Whitney criterion was tested. As a result of testing the hypothesis, the null theory of equality of samples was rejected, and a conclusion was made about the difference between samples in various combinations of a healthy speaker and a patient at different stages of treatment.

This approach currently does not show sufficient accuracy for its full implementation in the speech assessment methodology and needs additional refinement. However, the results obtained show the possibility of using this approach after additional research. As a direction for further work, it is proposed to study the optimal number and composition of participants in a group of healthy speakers (for example, only male healthy speakers for a male patient) or to determine the criterion for selecting an "ideal" healthy speaker from the group for comparison with the patient (for example, with the most matching voice characteristics).

Acknowledgments. This research was funded by the Ministry of Science and Higher Education of the Russian Federation within the framework of scientific projects carried out by teams of research laboratories of educational institutions of higher education subordinate to the Ministry of Science and Higher Education of the Russian Federation, project number FEWM-2020–0042.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Cancer statistics. https://spb.medsi.ru/articles/statistika-onkologicheskikh-zabolevaniy/. Accessed 24 May 2024
2. Lutsky L., Treger I.: Quality assessment in medical rehabilitation. Phys. Rehabil. Med. Med. Rehabil. **2**(1), 39–48 (2020). https://doi.org/10.36425/rehab19266
3. Novokhrestova, D.: Methods and algorithms of data analysis in assessing the quality of pronunciation of syllables in the process of speech rehabilitation. Thesis,. Doct. phylosophy (computer sci.). – Tomsk, p. 171 (2022)

4. Iordanishvili, A.K., et al.: Characteristics of relationship to disease in adult patients with chewing-speech apparatus pathology. Rossiyskiy stomatologicheskiy zhurnal. **20**(6), 309–314 (2016). https://doi.org/10.18821/1728-28022016;20(6):309-314
5. McFee, B., et al.: librosa: audio and music signal analysis in Python. In: SciPy, pp. 18–24 (2015). https://doi.org/10.25080/Majora-7b98e3ed-003
6. An Introduction to Dynamic Time Warping. https://builtin.com/data-science/dynamic-time-warping. Accessed 24 May 2024
7. ERP, R: Changing the distance with a real penalty (ERP). (r-project.org). Accessed 24 May 2024
8. MathWorks: EDR Edit distance on real signals. https://www.mathworks.com/help/signal/ref/edr.html. Accessed 24 May 2024
9. MSM Multi-scale Template Matching using Python and OpenCV. https://pyimagesearch.com/2015/01/26/multi-scale-template-matching-using-python-opencv/. Accessed 24 May 2024
10. LCSS: Longest Common Subsequence. https://www.geeksforgeeks.org/longest-common-subsequence-dp-4/. Accessed 24 May 2024
11. Euclidean Distance (Spatial Analyst). https://pro.arcgis.com/ru/pro-app/latest/tool-reference/spatial-analyst/euclidean-distance.htm. Accessed 24 May 2024
12. Rivero-Moreno, C., Escalante-Ramírez, B.: Seismic signal detection with time-frequency models. In: Proceedings of Third International Symposium on Time-Frequency and Time-Scale Analysis (TFTS-96), pp. 345–348 (1996). https://doi.org/10.1109/TFSA.1996.547484
13. Butyrsky, E.Y.: The Hilbert transform and its generalization. Sci. Instrum. **24**(4), 30–37 (2014). (in Russian)

A Comparative Study for Contextualized Spoken Answer Classification in German Medical Questionnaires

Philipp L. Harnisch(✉), Daniel Schuhmann, and Stefan Hillmann

Quality and Usability Lab, Technische Universität Berlin, Berlin, Germany
p.harnisch@tu-berlin.de

Abstract. This paper presents a study aimed at enhancing the classification accuracy of patients' spoken answer selections to German medical Patient-Reported Outcome Measures (PROM) questionnaires within a multimodal dialog system. We collected 1,737 speech data samples for training and evaluation through a lab experiment, employing textual priming as opposed to the visual priming utilized in prior research. For classification, we compare results from utilizing sentence embeddings against results from prompting various Large Language Models. We conduct a comparative analysis of approaches in terms of prediction performance, efficiency, hardware restraints, budget, inference time, and data privacy. Further, we investigate if adding the survey item text as context improves the classification. Results show the highest accuracy for `gpt-4` prompting, and indicate that including the questionnaire item text alongside user utterances is beneficial for LLM prompting. Additionally, we find significant positive correlations between accuracy and certain prompt characteristics.

Keywords: Text classification · Speech-to-text · German medical questionnaires · Large language models · Prompt engineering

1 Introduction

Patient-Reported Outcome Measures (PROM) surveys (see Fig. 1) are utilized in rehabilitation clinics and other healthcare settings to assess changes in patients' subjective health [5]. However, conventional survey methods, whether on paper or digitally, often yield low data quality and response rates in clinical settings [6,7]. Answering surveys with multimodal dialog systems with voice interfaces [4] are a promising approach to address this issue, because patients can answer in a more natural way that is less exhausting. It can help to assure that they choose the answer option that fits best to their situation by avoiding misunderstandings and fatigue from reading and ticking a document for a longer time. Therefore, we propose implementing a multimodal dialog system with a voice interface for answering PROM surveys. Depending on the preference, users should be able to tick answers on a tablet app, or interact with a voice interface

A. Karpov and V. Delić (Eds.): SPECOM 2024, LNAI 15299, pp. 377–391, 2024.
https://doi.org/10.1007/978-3-031-77961-9_28

Example PROM item description:
"Trotz meiner Beschwerden bin ich in der Lage schwierige Probleme zu lösen" ("Despite my complaints, I am able to solve difficult problems")

Answer options for this item:

1. "stimmt nicht" ("not true")
2. "stimmt wenig" ("true a little")
3. "stimmt mittelmäßig" ("true mediocre")
4. "stimmt ziemlich" ("pretty much true")
5. "stimmt sehr" ("very true")

Fig. 1. A health-related PROM item about the ability to solve difficult problems despite complaints.

to answer on survey items with speech. To increase the user experience, system users should not be limited to use the predefined answer option text exactly, but also free formulations that describe it (like synonyms or answers embedded in phrases or sentences). First, the implementation of such a voice interface requires deriving text from audio (Automatic Speech Recognition [ASR]). Second, it requires classifying the recognized text into the predefined answer options defined by the PROM survey (Natural Language Understanding [NLU]), which is the focus of this paper's investigations.

To train and evaluate, we collected 1,737 speech data samples from subjects in a lab study, all answering a PROM survey according to a pre-marked answer on the response scale. The aim of the experiment was to evaluate various classification approaches, using sentence embeddings (SE) [2] and Large Language Model (LLM) prompting [8]. Through this, we aimed to investigate our hypothesis that integrating survey item texts improves prediction performance.

2 Related Work

LLMs can be harnessed for many downstream tasks. For this, current research investigates the best methodologies to prompt LLMs to optimize LLM downstream task performances [9]. We want to apply this trend to our classification problem by adding the survey item text as additional context to the prompt.

In Harnisch and Hillmann [3], an empirical speech answer dataset is collected for a PROM questionnaire, and a sentence embedding classification approach is proposed and evaluated on the transcribed dataset. Participants were instructed to respond to the questionnaire items based not on their actual situation, but according to a randomly selected, visual priming on an ordered emoji-scale referring to different steps of happiness with the current health situation. The data collection approach and experimental setup from this prior work, are adopted and modified with changing the priming from visual to textual indication. In

addition, we add LLM prompting as classification approaches, and investigate the effect of added survey item text as additional context information.

3 Method

In this study, we want to evaluate classification approaches using different datasets. One of those datasets is created by our own experiment, which is also described in this method section.

3.1 PROM Questionnaire Items

In this paper, we investigate 92 standardized German PROM questionnaire items [1]. This is motivated by them being relevant to our project partner rehabilitation clinic, where field experiments are planned. All 92 PROM items share 13 different answer scales S_i (see Fig. 1), describing a spectrum from optimal to suboptimal health conditions. These scales vary in structure: 11 comprise five textual descriptions, one scale includes four textual descriptions, and another spans from 0 to 10 numerically. This selection of PROM questionnaire items is the same as used in the previous work [3].

3.2 Classification Datasets

We use three different datasets to evaluate various classification approaches. Firstly, we use the dataset $\mathcal{D}_V$ [3], which was collected through a visual priming experiment. Secondly, we create the dataset $\mathcal{D}_T$ by collecting data from a lab experiment with textual priming. The procedure of the experiment is described in Sect. 3.3. With those two datasets, we evaluate SE combined with linear classification layers, and LLM Prompting for classification of user utterances answering medical PROM survey items.

Thirdly, we use the same set of augmented text dataset $\mathcal{D}_A$ like Harnisch and Hillmann [3] for training of the approach SE'_A (see Table 2) to get insights on the helpfulness of augmented data in terms of performance. $\mathcal{D}_A$ was generated from 23 manually created templates combined with original answer option texts.

3.3 Empirical Collection of $\mathcal{D}_T$

To collect our dataset $\mathcal{D}_T$, we conducted an experiment with 19 participants who were on average 33.6 years old (10 male, 9 female). 6 participants had a mother tongue next to or other than German.

Our experiment approach resembles that of Harnisch and Hillmann [3]. PROM questionnaire items were presented to participants, and they were instructed to select an answer via voice according to a pre-defined priming on an answer scale. Therefore, we can assume that the ground truth of each recorded speech data point matches exactly the priming we made at this point in the experiment. This approach was chosen due to ethical concerns associated with

Table 1. Answer Option Scales.

Answ. Scale	Lang.	Answer options
S_1	DE	ausgezeichnet, sehr gut, gut, einigermaßen, schlecht
	EN	excellent, very good, good, moderately, bad
S_2	DE	vollständig, größtenteils, halbwegs, ein wenig, überhaupt nicht
	EN	fully, mostly, halfway, a little, not at all
S_3	DE	überhaupt nicht, ein wenig, mäßig, ziemlich, sehr
	EN	not at all, a little, moderate, quite, very much
S_4	DE	nie, selten, manchmal, oft, immer
	EN	never, rarely, sometimes, often, always
S_5	DE	keine Müdigkeit, schwach, mäßig, stark, sehr stark
	EN	no tiredness, mild, moderate, severe, very severe
S_6	DE	ohne jede Schwierigkeit, mit geringen Schwierigkeiten, mit einigen Schwierigkeiten, mit großen Schwierigkeiten, kann ich gar nicht
	EN	without any difficulty, with low difficulty, with some difficulty, with great difficulty, I can't at all
S_7	DE	überhaupt nicht, kaum, mäßig, ziemlich, kann ich gar nicht
	EN	not at all, barely, moderate, quite, I can't at all
S_8	DE	sehr schlecht, schlecht, mäßig, gut, sehr gut
	EN	very bad, bad, moderate, good, very good
S_9	DE	ich bin überhaupt nicht zuversichtlich, ich bin ein wenig zuversichtlich, ich bin etwas zuversichtlich, ich bin ziemlich zuversichtlich, ich bin sehr zuversichtlich
	EN	I'm not confident at all, I am a little confident, I am somewhat confident, I am somewhat confident, I am very confident
S_{10}	DE	kann ich gar nicht, mit großen Schwierigkeiten, mit einigen Schwierigkeiten, mit geringen Schwierigkeiten, ohne jede Schwierigkeiten
	EN	I can't at all, with great difficulty, with some difficulty, with low difficulty, without any difficulty
S_{11}	DE	nie, selten (einmal), manchmal (2-3 mal), oft (einmal pro Tag), sehr oft (mehrmals pro Tag)
	EN	never, rarely (once), sometimes (2-3 times), often (once a day), very often (several times a day)
S_{12}	DE	trifft gar nicht zu, trifft eher nicht zu , trifft eher zu, trifft völlig zu
	EN	not applicable at all, rather not applicable, rather applicable, fully applicable
S_{13}	DE	0 (keine Schmerzen), 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 (schlimmste vorstellbare Schmerzen)
	EN	0 (no pain), 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 (worst imaginable pain)

Table 2. Dataset information.

Dataset	$\mathcal{D}_A$ [3]	$\mathcal{D}_V$ [3]	$\mathcal{D}_T$
source	generated with templates	empirically collected	empirically collected
priming method	–	5-point happiness emoji scale	original text descriptions
n (data points)	1,467	1,793	1,737
words/n	5.79	16.34	11.07
characters/n	37.31	97.66	68.40

assessing sensitive health data of participants. Since the actual health data of the subjects is irrelevant to our study and arises ethical questions, priming was used to ensure that participants do not need to reveal their health status and sensitive information unnecessarily. In addition, this made possible to receive equally distributed data over the different answer option classes. This comes with the cost of potentially less realistic data, as the participants had to imagine themselves in a fictional health status.

The difference to the experiment of Harnisch and Hillmann [3] lies in the priming method. Instead of priming the subject by highlighting a specific emoji on an emoji answer scale, we use a verbal priming method to communicate the answer that the subject is supposed to choose (see Fig. 2). Our aim was to receive more realistic responses with this priming method change, as users may orientate themselves heavily on the displayed text options in real world usage.

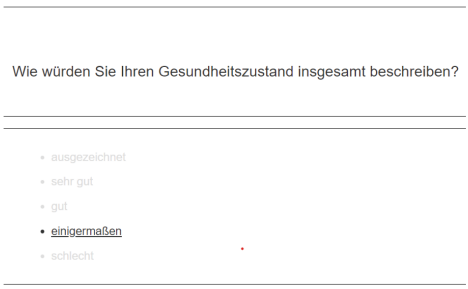


Fig. 2. Textual priming within lab study for $\mathcal{D}_T$. The survey item text is at the top, in the middle are all answer option texts, with one being highlighted.

3.4 Speech-to-Text Conversion

For further processing, spoken utterances from our experiment are transformed to text with a German Automatic-Text-Recognition (ASR) model [3], with a Word Error Rate (WER) of 17.1% and a Character Error Rate (CER) of 6.5% [3]. Despite this project related ASR model from a German startup is not openly

accessible for the reader, we decided on using it, because it will be the one used in the coming field experiments of our MIA-PROM research project.

3.5 Classification with SE and Contextual Linear Layers

This SE approach was already introduced in prior work [3]. It consists of the LaBSE [2] sentence embedding, followed by classification through a linear layer. For every survey answer option scale of the PROM survey (see Table 1), one linear layer is trained. Because the dialog system always knows which is the current question, it can derive which linear layer to use for current user utterance classification.

Before calculating embeddings, our SE approach combines questionnaire item text and user utterance, whilst SE$'$ only contains the user utterance (see Fig. 3). This enables us to investigate whether an integration of survey text for answer classification is beneficial or not.

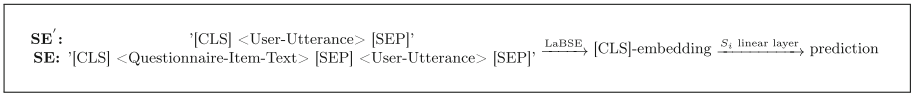


Fig. 3. Comparison of the sentence embedding (SE$^{(')}$) approaches for scale S_i answer classification. For SE the additional questionnaire item text is separated with the [SEP] token.

SE and SE$'$ are both trained solely on $\mathcal{D}_T$, but for SE$'_A$ we additionally used the augmented dataset $\mathcal{D}_A$. In this paper, all approaches denoted with an $'$ are without the use of the questionnaire item text, whilst all other approaches make use of it.

3.6 Classification with LLM Prompting

We wanted to tackle the task with a limited hardware setting (see Sect. 3.7) that could be realistic for real-world use in medical applications. Thus, we could not make experiments with larger models locally. We selected `llama3`[1], `llama2`[2], `Mistral`[3] and `gemma`[4] and run them with the `ollama` Python framework. With help of the OpenAI API, we also tested larger LLMs with `gpt-3.5-turbo`[5],

[1] https://ollama.com/library/llama3.
[2] https://ollama.com/library/llama2.
[3] https://ollama.com/library/mistral.
[4] https://ollama.com/library/gemma.
[5] https://platform.openai.com/docs/models/gpt-3-5-turbo.

and `gpt-4-turbo`[6] (denoted as "gpt-3.5" and "gpt-4" in the following for better readability). In Table 3, the number of parameters for gpt-4, and gpt-3.5 is estimated[7,8].

In our prompting template (see Fig. 4), we integrated a short task description, the survey item text, the user utterance, and all answer option texts with their ID.

```
Für einen Fragebogen-Gegenstand sollst du eine gesprochene Nutzer-Antwort
      der am besten passenden Antwort-Option zuordnen. Gebe dafür nur
     genau eine Antwort-Option (inklusive ID) in deiner Nachricht an und
     begründe anschließend deine Entscheidung.

Fragebogen-Gegenstand: '<Questionnaire-Item-Text>'

Nutzer-Aussage: '<User-Utterance>'

Antwort-Optionen: 'ID1: <First-Option-Text>, [...], ID<X>: <Last-Option-
     Text>'
```

Fig. 4. Prompt template for answer classification with questionnaire item text.

With `gpt-3.5-turbo`′, we excluded the item text to measure the contribution of it to the prediction performance. Find this template in Fig. 6 in the appendix.

3.7 Hardware

Training and inference of approaches was conducted on a laptop with a 4GB GDDR5 vRAM Quadro T2000 GPU.

3.8 Evaluation

Besides accuracy, we calculate the neighbor-accuracy, which also considers neighboring scale responses as correct, motivated by options being ordered. This is only used for evaluation purpose and never part of any training objective.

SE results stem from 5-fold cross-validation of the evaluation dataset, averaged over 10 different seeds.

We put emphasis on comparing approaches on $\mathcal{D}_T$, while comparing $\mathcal{D}_V$ and $\mathcal{D}_T$ with a subset of approaches.

[6] https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4.
[7] https://blog.wordbot.io/ai-artificial-intelligence/gpt-3-5-turbo-vs-gpt-4-whats-the-difference/.
[8] https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/.

4 Results

When comparing all tested approaches (see Table 3), we generally find SE and OpenAI LLM prompting approaches to outperform LLM prompting performed locally. By fair margin, `gpt-4` has the highest accuracy with 65.6%. We receive the highest neighbor-accuracy with our SE approach, with 82.3%. This stems from the low `gpt-4` performance on S_4 which makes up 508 of 1737 data points (see Tables 5 and 6 in the appendix).

Table 3. Comparison of $\mathcal{D}_T$ classification performance plus further relevant information of SE and LLM prompting approaches. Number of parameters for gpt-4, gpt-3.5 is estimated. Costs are Dollar per token.

$\mathcal{D}_T$ (n = 1737)	SE'_A	SE'	SE	gpt-4	gpt-3.5	gpt-3.5′	llama3	llama2	Mistral	gemma
Accuracy	0.590 ±0.010	0.587 ±0.010	0.573 ±0.010	**0.656**	0.575	0.549	0.543	0.329	0.554	0.492
Neighbor-acc.	0.787 ±0.008	0.808 ±0.007	**0.823** ±0.005	0.791	0.779	0.750	0.727	0.634	0.777	0.758
Params [B]	0.5	0.5	0.5	1,800 (?)	154 (?)	154 (?)	8	7	7	9
Time [s]	**0.4**	**0.4**	**0.4**	6.7	2.3	2.3	82.2	28.5	34.4	35.8
Disk [GB]	1.9	1.9	1.9	–	–	–	4.7	3.8	4.1	5.0
Cost [$]	–	–	–	0.0060	0.0002	0.0002	–	–	–	–
Data privacy	**Yes**	**Yes**	**Yes**	No	No	No	**Yes**	**Yes**	**Yes**	**Yes**

Table 4. Comparison of $\mathcal{D}_V$ classification performance of SE and LLM prompting approaches.

$\mathcal{D}_V$ **(n = 1793)**	SE'_A	SE'	SE	gpt-4	gpt-3.5	gpt-3.5′	gemma
Accuracy	0.511 ±0.012	0.505 ±0.008	0.495 ±0.010	**0.562**	0.485	0.468	0.405
Neighbor-acc.	0.776 ±0.009	0.789 ±0.010	**0.799** ±0.008	0.793	0.784	0.743	0.747

For `gpt-3.5`, the integration of item text results in a performance increase of +2.6% for accuracy and +2.9% for neighbor-accuracy. For SE, this results in a performance change of –1.4% for accuracy and +1.5% for neighbor-accuracy.

In relation to their parameter size, the SE approaches are most efficient, leading to the lowest inference time and (local) disk space requirement. In contrast to SE or local LLM prompting approaches, the OpenAI API costs money, and data has to be sent to a remote server.

4.1 Influence of Item Text and User Utterance Length on Performance

From further analysis, we find significant correlations for `gpt-3.5` and parts of the prompt. Item text length correlates positively with accuracy (15%) and neighbor-accuracy (20%). Item text being formulated as a question correlates positively with accuracy (11%) and neighbor-accuracy (19%). Further, we find a positive correlation for the length of user utterance and neighbor-accuracy of 9%.

4.2 Comparison of $\mathcal{D}_T$ and $\mathcal{D}_V$

Comparing evaluation results of $\mathcal{D}_T$ and $\mathcal{D}_V$ (see Tables 3 and 4), we can observe similar relative differences between the different approaches. But in absolute numbers, all accuracy scores of all approaches listed in Table 4 are 8.5% lower for $\mathcal{D}_V$ than for $\mathcal{D}_T$ in average.

5 Discussion

In this study, we wanted to investigate whether integrating survey item texts improves prediction performance of different text classification approaches. Higher scores for `gpt-3.5` than `gpt-3.5'` indicate that this hypothesis holds true. However, for the sentence embedding classification approach, we find lower accuracy with this additional context (despite higher neighbor-accuracy). This could be due to noise in the data, or that the survey item text integration may only be beneficial for the LLM classification approach.

The results of the present study show that `gpt-4` has the highest accuracy. However, unlike SE or local LLM prompting approaches, the OpenAI API costs money and data is sent to a remote server. Therefore, working with sensitive health data leads to severe ethical questions in regard to data privacy when using OpenAI LLMs.

The higher performance metrics for $\mathcal{D}_T$ in contrast to those for $\mathcal{D}_V$ indicates that $\mathcal{D}_T$ is more realistic and less noisy than data from prior work [3]. This can be explained by the main difference of the two datasets, the priming method. Priming verbally with a word on a verbal scale instead of an emoji on an emoji scale should lead to more precise responses, because the same emojis were used for different scales and left more space for interpretation.

The models `llama2` and `Mistral`, that we selected, both often use English within their answer and reasoning, indicating possible difficulties with the German language. `gpt-4` and `Mistral` sometimes refuse to answer when they are confident that the user utterance is not precise enough or does not give enough information. `gemma`'s reasoning is sometimes contradictory (an example can be found in the appendix – Fig. 5).

5.1 Limitations

Our lab experiment, in which participants were primed to provide formulations for answer options highlighted by us, poses a notable limitation, as given answers may vary from real-world usage of participants responding to voice interface dialog systems with their own health status to fill out medical surveys. Furthermore, the evaluation data derived from the experiment exhibits noise, likely provoked by participants being irritated from the varied formulation of item descriptions – some positive and others negative.

The results may be influenced by the relatively small number of participants, which is not representative of the general population.

Moreover, this experiment was conducted by assessing 92 standardized German PROM questionnaire items, because they are relevant to our project partner rehabilitation clinic. Therefore, the results do not necessarily apply to other (health) questionnaires.

A significant constraint arises for our discussion about hardware restraints, which may vary across different scenarios. Deploying larger LLMs could enhance classification accuracy without reliance on external APIs. Because we put effort to optimize a selection of suitable classification approaches for the integration of questionnaire item texts for survey answer classification, we had to limit the amount of models tested.

6 Conclusion

To conclude, we find `gpt-4-turbo` to have the highest accuracy, but this approach has severe drawbacks with its high inference time, costs, and challenges with data privacy. Therefore, our SE approach is more suitable to classify user utterances, because it has the highest neighbor-accuracy overall, and is also fast enough to give user feedback. Still, the SE approaches need to learn every distinct answer option scale, in contrast to the LLM prompting variants. This requires new pre-training if surveys are added to the system with unlearned answer option scales.

The integration of survey item texts within classification approaches is beneficial for LLM prompting, especially for longer item texts being formulated as a question. But we could not find clear results for the SE approach.

If the response of a participant does not match the priming at all, it is most likely because of confusion or a misunderstanding. These occurrences should be removed, as they have no meaningful contribution. We plan on cleaning the dataset $\mathcal{D}_T$ in the future in order to reduce noise of the data. If this is successful, we will make the dataset publicly available afterwards.

6.1 Ethical Considerations

The participants were informed about the procedure of the experiment and had the option to quit at any time if they wished to do so. The PROM questionnaire

collects sensitive and personal health data. However, because the actual health data of the participants was not relevant for this study, they were instructed to answer according to a highlighted word on the answer scale rather than how they actually felt.

Acknowledgments. The presented work has been funded by the Federal Ministry of Education and Research (Germany) under grant no. 16SV9018 for the joint research project MIA-PROM in the research program Interactive Technologies for health and Quality of Life.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

A Appendices

This appendix lists tables and figures with further details to our main research description.

A.1 LLM Prompting

Figure 5 shows an example response of `gemma` with unlogical reasoning. Figure 6 displays the German prompting template without survey item text. The English translation of templates with and without item text can be found in Fig. 7 and Fig. 8.

A.2 Performance per Answer Option Scale

Table 5 and Table 6 contain performance metrics of our evaluation experiments separately per answer option scale, on $\mathcal{D}_T$ and $\mathcal{D}_V$ respectively.

```
Die Nutzer-Aussage besagt, dass sie keine Probleme mit der Nachfolgung
    von Aktivitäten hat. Dies ist mit der Antwort-Option "manchmal" am
    besten in Einklang [...]
```

```
The user statement says that they have no problems with tracking
    activities. This is most consistent with the 'sometimes' answer
    option [...] (translated from German)
```

Fig. 5. Example for unlogical reasoning of gemma.

```
Für einen Fragebogen-Gegenstand sollst du eine gesprochene Nutzer-Antwort
      der am besten passenden Antwort-Option zuordnen. Gebe dafür nur
     genau eine Antwort-Option (inklusive ID) in deiner Nachricht an und
     begründe anschließend deine Entscheidung.

Nutzer-Aussage: '<User-Utterance>'

Antwort-Optionen: 'ID1: <First-Option-Text>, [...], ID<X>: <Last-Option-
     Text>'
```

Fig. 6. Prompt template for answer classification without questionnaire item text.

```
For a questionnaire item, you should assign a spoken user response to the
      most appropriate answer option. Provide only one answer option (
     including ID) in your message and then justify your decision.

Questionnaire Item: '<Questionnaire-Item-Text>'

User Statement: '<User-Utterance>'

Answer Options: 'ID1: <First-Option-Text>, [...], ID<X>: <Last-Option-
     Text>'
```

Fig. 7. Translated prompt template for answer classification with questionnaire item text.

```
For a questionnaire item, you should assign a spoken user response to the
      most appropriate answer option. Provide only one answer option (
     including ID) in your message and then justify your decision.

User Statement: '<User-Utterance>'

Answer Options: 'ID1: <First-Option-Text>, [...], ID<X>: <Last-Option-
     Text>'
```

Fig. 8. Translated prompt template for answer classification without questionnaire item text.

Table 5. Comparison of $\mathcal{D}_T$ classification performance of SE and LLM prompting approaches, per answer scale.

$\mathcal{D}_T$ (n = 1737)		SE'_A	SE'	SE	gpt-4	gpt-3.5	gpt-3.5′	llama3	llama2	Mistral	gemma
S_1 (n=114)	Accuracy	0.630 ±0.033	0.538 ±0.038	0.447 ±0.025	**0.772**	0.658	0.640	0.623	0.228	0.719	0.465
	Neighbor-acc.	0.892 ±0.025	0.844 ±0.028	0.791 ±0.022	0.912	**0.921**	0.886	0.851	0.605	**0.921**	0.877
S_2 (n=57)	Accuracy	0.640 ±0.050	0.502 ±0.049	0.412 ±0.046	**0.807**	0.702	0.719	0.667	0.281	0.719	0.526
	Neighbor-acc.	0.785 ±0.032	0.777 ±0.028	0.715 ±0.072	**0.965**	**0.965**	0.947	0.825	0.614	0.947	0.772
S_3 (n=435)	Accuracy	0.642 ±0.022	0.632 ±0.020	0.606 ±0.019	**0.715**	0.651	0.634	0.602	0.366	0.614	0.623
	Neighbor-acc.	0.854 ±0.008	0.850 ±0.018	0.851 ±0.014	0.876	0.867	0.862	0.811	0.708	0.841	**0.885**
S_4 (n=508)	Accuracy	0.456 ±0.020	0.584 ±0.015	**0.655** ±0.015	0.427	0.364	0.346	0.380	0.301	0.348	0.317
	Neighbor-acc.	0.567 ±0.026	0.688 ±0.021	**0.775** ±0.013	0.480	0.467	0.453	0.484	0.522	0.484	0.533
S_5 (n=19)	Accuracy	0.555 ±0.057	0.270 ±0.046	0.245 ±0.035	**0.737**	0.684	0.579	0.526	0.368	0.579	0.632
	Neighbor-acc.	**0.925** ±0.040	0.710 ±0.097	0.695 ±0.069	0.895	0.895	0.842	0.737	0.842	0.895	0.842
S_6 (n=74)	Accuracy	0.672 ±0.043	0.627 ±0.032	0.555 ±0.057	**0.797**	0.689	0.757	0.676	0.419	0.743	0.541
	Neighbor-acc.	0.899 ±0.029	0.903 ±0.026	0.889 ±0.039	**0.986**	0.973	0.959	0.892	0.676	0.959	0.878
S_7 (n=94)	Accuracy	0.625 ±0.048	0.546 ±0.072	0.413 ±0.050	**0.830**	0.660	0.564	0.521	0.479	0.596	0.670
	Neighbor-acc.	0.854 ±0.034	0.816 ±0.047	0.800 ±0.040	0.915	0.894	0.787	0.681	0.660	0.894	**0.926**
S_8 (n=19)	Accuracy	0.465 ±0.105	0.170 ±0.078	0.190 ±0.062	**0.789**	0.684	0.684	0.737	0.526	0.632	0.526
	Neighbor-acc.	0.975 ±0.034	0.705 ±0.101	0.720 ±0.078	**1.000**	**1.000**	0.947	**1.000**	0.947	0.947	0.947
S_9 (n=152)	Accuracy	0.643 ±0.029	0.598 ±0.030	0.523 ±0.034	**0.730**	0.684	0.618	0.664	0.283	0.651	0.553
	Neighbor-acc.	0.917 ±0.020	0.926 ±0.019	0.888 ±0.020	**0.928**	0.908	0.908	**0.928**	0.599	**0.928**	0.829
S_{10} (n=76)	Accuracy	0.732 ±0.037	0.642 ±0.049	0.605 ±0.045	**0.855**	0.750	0.724	0.724	0.303	0.671	0.645
	Neighbor-acc.	0.876 ±0.029	0.879 ±0.034	0.910 ±0.047	0.961	0.947	0.961	0.934	0.750	**0.974**	0.868
S_{11} (n=95)	Accuracy	0.683 ±0.034	0.649 ±0.033	0.562 ±0.043	**0.832**	0.695	0.695	0.600	0.316	0.642	0.558
	Neighbor-acc.	0.913 ±0.023	0.899 ±0.023	0.852 ±0.039	**0.989**	0.968	0.884	0.779	0.600	0.916	0.811
S_{12} (n=76)	Accuracy	0.720 ±0.044	0.660 ±0.048	0.742 ±0.062	**0.750**	0.645	0.526	0.579	0.368	0.645	0.368
	Neighbor-acc.	0.950 ±0.022	0.954 ±0.025	**0.974** ±0.018	**0.974**	**0.974**	0.763	0.803	0.829	0.961	0.671
S_{13} (n=18)	Accuracy	0.125 ±0.072	0.060 ±0.054	**0.130** ±0.087	0.000	0.000	0.000	0.000	0.000	0.056	0.000
	Neighbor-acc.	0.505 ±0.085	0.375 ±0.068	0.505 ±0.088	**0.722**	0.611	0.556	0.556	0.611	**0.722**	0.556

Table 6. Comparison of $\mathcal{D}_V$ classification performance of SE and LLM prompting approaches, per answer scale.

$\mathcal{D}_V$ (n = 1793)		SE'_A	SE'	SE	gpt-4	gpt-3.5	gpt-3.5′	gemma
S_1 (n=119)	Accuracy	0.616 ±0.023	0.578 ±0.035	0.557 ±0.038	**0.664**	0.605	0.580	0.529
	Neighbor-acc.	0.923 ±0.019	0.913 ±0.020	0.902 ±0.019	**1.000**	**1.000**	0.958	0.882
S_2 (n=60)	Accuracy	0.478 ±0.051	0.483 ±0.048	0.357 ±0.055	**0.700**	**0.700**	0.667	0.417
	Neighbor-acc.	0.845 ±0.029	0.787 ±0.035	0.710 ±0.033	0.950	**0.983**	0.950	0.817
S_3 (n=457)	Accuracy	0.573 ±0.017	0.567 ±0.017	0.556 ±0.019	**0.650**	0.527	0.501	0.551
	Neighbor-acc.	0.826 ±0.019	0.834 ±0.020	0.836 ±0.012	**0.902**	0.867	0.803	0.877
S_4 (n=539)	Accuracy	0.411 ±0.029	0.448 ±0.019	**0.502** ±0.015	0.345	0.291	0.321	0.232
	Neighbor-acc.	0.642 ±0.018	0.688 ±0.025	**0.751** ±0.014	0.482	0.501	0.532	0.551
S_5 (n=20)	Accuracy	0.355 ±0.088	0.210 ±0.109	0.270 ±0.081	**0.800**	0.600	0.450	0.550
	Neighbor-acc.	0.715 ±0.067	0.680 ±0.081	0.580 ±0.110	**0.900**	0.850	0.800	**0.900**
S_6 (n=80)	Accuracy	0.619 ±0.071	0.604 ±0.068	0.546 ±0.032	**0.738**	0.688	0.675	0.388
	Neighbor-acc.	0.905 ±0.031	0.924 ±0.025	0.881 ±0.022	**0.988**	0.975	0.963	0.875
S_7 (n=100)	Accuracy	0.573 ±0.059	0.525 ±0.046	0.467 ±0.047	**0.650**	0.520	0.390	0.490
	Neighbor-acc.	0.883 ±0.026	0.887 ±0.029	0.867 ±0.035	**0.980**	0.950	0.620	0.840
S_8 (n=20)	Accuracy	0.585 ±0.071	0.210 ±0.073	0.275 ±0.087	**0.850**	0.750	0.650	0.600
	Neighbor-acc.	0.980 ±0.024	0.735 ±0.100	0.885 ±0.087	**1.000**	**1.000**	0.950	**1.000**
S_9 (n=144)	Accuracy	0.540 ±0.045	0.520 ±0.032	0.454 ±0.044	**0.562**	0.549	0.479	0.465
	Neighbor-acc.	0.831 ±0.028	0.837 ±0.026	0.826 ±0.033	**0.903**	0.896	0.882	0.799
S_{10} (n=72)	Accuracy	0.616 ±0.042	0.545 ±0.033	0.533 ±0.045	0.722	0.667	**0.764**	0.514
	Neighbor-acc.	0.845 ±0.040	0.843 ±0.036	0.824 ±0.038	**0.986**	0.958	0.944	0.861
S_{11} (n = 90)	Accuracy	0.452 ±0.046	0.428 ±0.035	0.399 ±0.036	**0.589**	0.522	0.544	0.267
	Neighbor-acc.	0.721 ±0.039	0.766 ±0.039	0.751 ±0.036	**0.922**	0.900	0.789	0.722
S_{12} (n = 72)	Accuracy	0.532 ±0.053	0.460 ±0.031	0.428 ±0.044	**0.750**	0.625	0.486	0.347
	Neighbor-acc.	0.845 ±0.031	0.833 ±0.026	0.821 ±0.021	0.889	**0.917**	0.819	0.639
S_{13} (n = 20)	Accuracy	0.200 ±0.089	0.195 ±0.104	0.115 ±0.059	**0.350**	0.200	0.250	0.250
	Neighbor-acc.	0.245 ±0.104	0.235 ±0.087	0.180 ±0.040	**0.500**	0.350	0.400	0.400

References

1. Cella, D., et al.: PROMIS ®adult health profiles: efficient short-form measures of seven health domains. Value Health **22**, 537–544 (2019)
2. Feng, F., Yang, Y., Cer, D., Arivazhagan, N., Wang, W.: Language-agnostic BERT sentence embedding. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 878–891, Dublin, Ireland. Association for Computational Linguistics (2022)
3. Harnisch, PL., Hillmann, S.: Empirical evaluation of asr and nlu in a multimodal dia- logue system for survey answering. In Elektronische Sprachsignalverarbeitung 2024, Tagungsband der 35. Konferenz, Regensburg, 6.-8. März 2024, pp. 211–218 (2024)
4. Hillmann, S., et al.: Multimodal interactive assistance for the digital collection of patient-reported outcome measures. In: The Digitalization of Healthcare for Older Adults, Berlin, vol. 6 (2024) [In press]
5. Kluzek, S., Dean, B., Wartolowska, K.A.: Patient-reported outcome measures (PROMs) as proof of treatment efficacy. BMJ Evi. Based Med. **27**(3), 153–155 (2022)
6. Köhn, S., et al.: Predicting non-response in patient-reported outcome measures: results from the Swiss quality assurance programme in cardiac inpatient rehabilitation. Inter. J. Quality Health Care, **34**(4), mzac093 (2022)
7. Kutschar, P., Weichbold, M., Osterbrink, J.: Effects of age and cognitive function on data quality of standardized surveys in nursing home populations. BMC Geriatr. **19**(1), 244 (2019)
8. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput. Surv., **55**(9) (2023)
9. Sahoo, P., Singh, A.K., Saha, S., Jain, V., Mondal, S., Chadha, A.: A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications. arXiv preprint arXiv: 2402.07927v1 (2024)
10. Zhang, Z., Yu, T., Zhao, H., Xie, K., Yao, L., Li, S.: Exploring soft prompt initialization strategy for few-shot continual text classification. In ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 12106–12110 (2024)

Author Index

A. Karpov and V. Delić (Eds.): SPECOM 2024, LNAI 15299, pp. 393–395, 2025.
https://doi.org/10.1007/978-3-031-77961-9