Syllable Stress Detection to Evaluate Pathological Speech and L2 Language Level Pronunciation

Juan Camilo Vásquez-Correa(✉), Haritz Arzelus, and Aitor Álvarez

Vicomtech Foundation, Basque Research and Technology Alliance (BRTA), Donostia - San Sebastián 20009, Spain
{jcvasquez,harzelus,aalvarez}@vicomtech.org
https://www.vicomtech.org

Abstract. Syllable stress is a major key aspect of speech frequently altered by neurological impairments and varying levels of non-native language proficiency. This paper presents a novel syllable stress detection pipeline designed to evaluate pathological speech and assess L2 learner pronunciation. Separate models for Spanish and English were trained using Common Voice and Librispeech corpora. The proposed architecture extracts multidimensional syllable-level acoustic representations, combining prosodic, spectral, and articulatory features. These representations are processed through a 1D-convolutional projection and a Transformer encoder to capture inter-syllabic dependencies, followed by a time-distributed multi-layer perceptron for per-syllable stress prediction. In-domain evaluations demonstrated overall accuracies of 97.5% for Spanish and 93.9% for English. Furthermore, the model's practical utility was validated through benchmark evaluations on out-of-domain datasets, spanning L2 Spanish learners and clinical populations with Aphasia, Alzheimer's, and Parkinson's diseases. The model successfully captured deviations in stress patterns, with detection accuracy reflecting both Aphasia severity and Spanish L2 proficiency levels. These findings suggest that automated syllable stress detection serves as a viable, objective tool for clinical speech assessment and language learning tracking.

Keywords: Syllable Stress Detection · Pathological Speech Processing · L2 level Assessment

1 Introduction

Syllable stress involves the emphasis or prominence placed on a specific syllable within a word [7]. Proper stress placement is a fundamental mechanism in stress-timed languages such as English and Spanish. It influences word meaning and helps to direct a listener's attention to critical information within an utterance [15]. Incorrect stress on syllables can alter the intended meaning of a word, leading to confusion and reduced intelligibility in human-human and human-machine interactions [1]. Deviations in stress patterns are prevalent across two

E. Nöth et al. (Eds.): TSD 2026, LNAI 16940, pp. 277–288, 2027.
https://doi.org/10.1007/978-3-032-37249-9_23

distinct populations: First, prosodic disturbances, including a reduced ability to appropriately vary stress and accent, are considered a perceptual biomarker of motor speech disorders and neurological impairments such as dysarthria [17]. Additionally, second language (L2) learners frequently misplace stress due to the influence of rhythm and stress patterns carried over from their native languages [1,15]. Given these challenges, automated syllable stress detection has emerged as a crucial component for providing localized pronunciation feedback in computer assisted language learning systems, as well as an objective tool to support the clinical assessment of pathological speech.

Traditionally, automatic stress detection has been treated as a binary classification problem relying on prosody and sonority-based features [11,17,22]. In particular, related studies have shown that stressed syllables differ mainly in fundamental frequency (f_0), intensity, and duration [11]. Early systems typically used classifiers such as Support Vector Machines or Gaussian Mixture Models operating on these handcrafted acoustic features [17]. Subsequent advancements recognized that stress is highly dependent on phonetic context and syllabic structure. Consequently, researchers had incorporated contextual rules and sonority-based features to enhance detection accuracy [11,22].

In recent years, deep learning architectures have advanced the state-of-the-art by capturing complex, non-linear relationships across acoustic features. Multi-distribution Deep Neural Networks (DNNs) integrating lexical and syntactic information, and self-supervised representations like Wav2vec 2.0 have demonstrated robust feature extraction capabilities [1,14,15]. Furthermore, due to stress is a suprasegmental phenomenon, meaning the perception of prominence depends on the surrounding syllables, recent studies have shifted from sequence-independent models to sequence-dependent architectures. Models like Long Short-Term Memory networks and Transformers have been successfully applied to capture these inter-syllabic dependencies at the word or utterance level [1,15].

Despite these architectural advancements, existing stress detection models are typically evaluated in narrow, isolated scenarios, focusing either exclusively on specific L2 cohorts or on single neurological conditions. Building upon the necessity of contextual acoustic modeling, this paper proposes a lexical stress detection pipeline designed to bridge this gap. The proposed approach combines the computation of 21 prosodic features with 85 spectral and articulatory markers, and a deep learning model to capture multidimensional inter-syllabic dependencies. To the best of our knowledge, this is the first study to conduct a large-scale, cross-domain evaluation of automated stress detection across such a broad spectrum of pathological and non-native speech benchmarks. We evaluate the proposed models against in-domain benchmarks, and subsequently validate their practical clinical and educational utility on diverse, out-of-domain datasets covering L2 learners and patients affected by Aphasia, Alzheimer's disease (AD), and Parkinson's disease (PD). Despite the individual neural components are established, the novelty lies in the large-scale, cross-domain evaluation of these techniques across multiple pathological and non-native speech benchmarks.

2 Syllable Stress Detection

The proposed syllable stress detection pipeline is shown in Fig. 1. It consists of three main stages, which are described in detail in the following subsections.

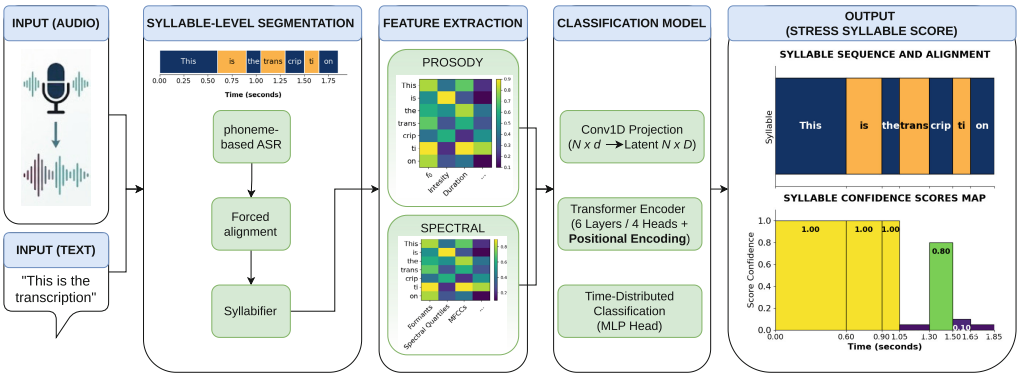

Fig. 1. Pipeline for syllable-level stress detection. The process consists of three main stages: (1) Syllable-level Segmentation: Forced alignment is performed between raw audio and its transcription, using a phoneme-based ASR model to define temporal boundaries. (2) Feature Extraction: Within each boundary, two sets of feature vectors are computed, capturing prosodic and spectral components. (3) Transformer-based Classification: Feature sequences are projected via a 1D-Convolutional layer, enriched with positional embeddings, and processed through a Transformer Encoder to model inter-syllabic prosodic dependencies. The final time-distributed multi-layer perceptron (MLP) head outputs a stress probability score for each syllable in the sequence.

2.1 Syllable-Level Segmentation

The syllable segmentation process was performed independently for each language, considering specific constraints to ensure accuracy in defining the temporal boundaries of the syllables. First, phonetic transcriptions for every text were generated. For Spanish, a rule-based automatic phonetic transcription system, available online[1] was used, while for English, a Grapheme-to-Phoneme (G2P) model was considered. This model was trained using Sequitur-G2P[2] [3], using a CMU dictionary[3] as the reference lexicon. The second step involved phone-level forced alignment, where each phone in the transcription was aligned to the audio. The process was performed using Kaldi [20] with a beam size of 1 and a retry beam of 2 to discard non-literal transcriptions. In addition, we considered a frame subsampling factor of 1 to obtain more accurate start times and durations for each phone. The third step consisted of syllabifying each word in the transcriptions. For Spanish, syllabification was performed using Silabeador

[1] https://github.com/xavierlopez/automatic-phonetic-transcriptor.
[2] https://github.com/sequitur-g2p/sequitur-g2p.
[3] http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

TIP [8], which enables the automatic division of phonetic transcriptions into syllabic units. For English, two tools were considered: (1) Syllabificator[4] to convert each word into its syllable representations, and (2) syllabify[5] to match each aligned phone into the syllable sequences. In cases where syllables could not be extracted from a transcription or when phoneme-syllable alignment could not be determined with certainty, the corresponding audio was discarded and excluded for the training process. Finally, the last step comprised generating TextGrid files with syllable segments for each audio, for further analysis.

2.2 Feature Extraction

The stress detection models were trained with two groups of features computed for each segmented syllable within each word. The first one comprised 21 syllable-level prosodic features obtained from pitch and intensity contours, and the sylable duration. The process begins by computing the first and second derivatives of the fundamental frequency (f_0) and intensity contours, to capture dynamic changes in speech. f_0 is extracted using the CREPE model [10], while intensity was computed following the Praat algorithm[6]. For each syllable, we extracted statistical measures (mean, max, min) and relative measures (deviations from global medians) for pitch and energy, following [17]. Additionally, we computed specialized prosodic markers from [11], including V_{dur} (log-normalized duration), V_{loud} (maximum loudness), and V_{pitch} (initial vs. final semitone change), which provided a multidimensional representation of the rhythmic and melodic characteristics of the speech signal.

The second feature set comprised 85 spectral and articulation features computed per syllable, including the spectral centroid, the first, second, and third quartiles of the spectrum, the mean frequencies of the first three formants, and the mean and standard deviation of 13 Mel Frequency Cepstral Coefficients (MFCCs) along with their first and second derivatives.

2.3 Classification Network

We implemented a Transformer-based stress detection model to process sequences of syllable-level acoustic features. The architecture comprises a multistage approach to map input features into a stress-probability space. The first part of the model is a feature projection, where the input features are processed through a 1D convolution layer with a kernel size of 1. This acts as a learnable linear projection that maps the raw acoustic dimensions into a higher-dimensional latent space. This stage is followed by batch normalization and dropout to stabilize training and mitigate over-fitting. The second part of the model consists of a sequential encoding mechanism to capture the rhythmic and temporal dependencies between syllables. We employed a Transformer encoder

[4] https://github.com/guslatho/syllabificator.
[5] https://github.com/cainesap/syllabify.
[6] https://www.fon.hum.uva.nl/praat/manual/Sound__Get_intensity__dB_.html.

consisting of six layers with four attention heads each. We used pre-layer Normalization (norm-first) and GELU activation functions to improve gradient flow. Learned positional embeddings were added to the projected sequence to provide the self-attention mechanism with relative temporal context. The final part of the model corresponds to a time-distributed classification, where the final hidden states are passed through a multi-layer perceptron head. This classifier applies layer normalization followed by a bottleneck linear layer and a final projection to a single scalar per time step. This configuration allows the model to perform per-syllable stress prediction across the entire temporal sequence simultaneously.

3 Data

3.1 Training Data

Separate syllable stress detection models were trained for Spanish and English languages, respectively. The model for Spanish was trained with audios selected from Common Voice [2], while Librispeech [19] (train-other partition) was used to train the English model. For Librispeech, 193 h of audio (38% of the corpus) were excluded from the training set because a reliable correspondence between phonetic transcriptions and syllable-level boundaries could not be established. Table 1 summarizes the main aspects of the final corpora used for training the models in the two target languages. Both corpora were randomly split into training (80%), development (10%), and test (10%) sets, in a speaker independent fashion.

Table 1. Key statistics of the training datasets, detailing the audio duration, file counts, and total words for both target languages.

Language	Spanish	English
Corpus	Common Voice	Librispeech
Audio files	289 209	88 613
Audio duration (hours)	411	304
# Words	1 507 813	696 362

3.2 Benchmark Evaluation

The list of the following corpora were used to externally validate the performance of the trained models. They were also considered to validate the suitability of syllable stress detection as a tool to evaluate the pronunciation capabilities of speakers in constrained scenarios, focused on two applications: (1) evaluating the pronunciation deficits of patients with neurological impairments such as Aphasia, AD, or PD with respect to Healthy Control (HC) subjects, and (2) assessing the proficiency level of students learning a second language. Table 2 summarizes the main aspects of the benchmark test corpora. We relied on the internal control

Table 2. Key aspects of the different benchmark test corpora considered to evaluate the impact of syllable stress detection on pathological and L2 language pronunciation speech. **Duration** of the corpus expressed in hours. The **# of syllables per word** expressed as median (standard deviation).

Corpus	Speakers	Duration	Words	Syll. per word
AphasiaBank-ES [13]	4 Patients	1.2	1571	2 (0.56)
AphasiaBank-EN [13]	277 HC, 386 Patients	141	58508	2 (0.59)
Ivanova [9]	197 HC, 90 MCI, 74 AD	3.8	10819	2 (0.77)
Neurovoz [16]	55 HC, 53 PD	2.3	9237	2 (0.61)
ADReSS-M [12]	122 HC, 114 AD	5.0	5241	2 (0.43)
SPLLOC [18]	80	21.8	62276	2 (0.71)

groups provided by these standard datasets, thus demographic matching was limited to what the original corpus creators provided.

AphasiaBank [13]**:** This corpus includes interviews between patients diagnosed with Aphasia and clinicians. This study considered the English and Spanish version of the corpus. The English partition covers 141 h of speech material from 663 speakers (277 HC, and 386 patients), while the Spanish partition includes only 1.2 h from 4 patients. The Aphasia severity of patients is labeled with the Western Aphasia Battery (WAB) scale, which ranged from 0 to 100. Lower scores meant a higher degree of Aphasia severity. These scores served as a threshold to classify the diagnosed patients into four aphasic levels, including mild (WAB $>$ 75), moderate ($50 <$ WAB ≤ 75), severe ($25 <$ WAB ≤ 50) and very severe ($0 <$ WAB ≤ 25).

Ivanova [9]**:** The DementiaBank Spanish Ivanova corpus was created to evaluate speech impairments in patients affected by Mild Cognitive Impairment (MCI) and AD. It contains utterances from 74 AD, 90 MCI, and 197 HC subjects reading the first two sentences of Don Quixote by Cervantes. The 281 utterances from equal number of speakers correspond to 5 h of audio. The age of the participants ranges from 50 to 96 years old. The Mini Mental State Examination (MMSE) was administered to all subjects to evaluate their cognitive capabilities. The average MMSE score is 28 ($\sigma = 1.9$) for HC subjects, 24 ($\sigma = 4$) for MCI patients, and 20 ($\sigma = 5$) for AD patients.

Neurovoz [16]: This corpus was designed to evaluate speech and language impairments in PD patients. It contains utterances from 108 participants, including 53 PD patients and 55 HC subjects, who perform several speech tasks. The read sentences and spontaneous speech utterances cover a total of 2.3 h of data. The neurological state of patients was evaluated with the Hoehn&Yahr stadium (from 1 to 5). The majority of patients are in stadium 2 and 3. Similar to Ivanova and Aphasiabank corpora, these data are used to evaluate the impact in the accuracy of the syllable stress detection in this population and whether there are differences between results for HC and PD patients.

ADReSS-M [12]**:** These data corresponded to the training corpus from the ICASSP 2023 challenge on Multilingual Alzheimer's Dementia Recognition through Spontaneous Speech [12]. The corpus consists of 5 h of English spontaneous speech samples corresponding to picture descriptions produced by 114 HC subjects and 122 AD patients, who were asked to describe the Cookie Theft picture from the Boston Diagnostic Aphasia Examination test. The cognitive level of the subjects was labeled with the MMSE scale. The average MMSE score is 28.9 ($\sigma = 1.1$) for HC subjects and 17.8 ($\sigma = 5.5$) for AD patients.

The Spanish Learner Language Oral Corpora (SPLLOC) [18]**:** This is a corpus of L2 Spanish that includes 22 h of spoken Spanish produced by 60 instructed learners with English as their L1. Those non native speakers are divided into 3 equally distributed groups of beginners (age = 13 - 14 years old and A2 Spanish level), intermediate (age = 17 - 18 years old and B1-B2 Spanish level), and advanced (age = 21 - 22 years old and C1-C2 Spanish level). The corpus also includes samples of spoken Spanish produced by native speakers of similar ages to the L2 learners (5 at each age level), performing the same range of tasks. The subjects performed a set of specially designed elicitation tasks, including storytelling, picture description, discussion and individual interviews. SPLLOC is considered here to evaluate the capabilities of the speakers in each group to accurately mark the stress in each of the pronounced words.

4 Results and Discussion

This section presents the evaluation of the trained models for syllable stress detection. The overall in-domain performance is shown in Table 3. The evaluation included the performance of models trained with prosody and spectral feature sets, and their fusion, at feature level. The reported accuracy reflects the proportion of words where the stress location was correctly identified. Because the system evaluates stress placement as a word-level multiclass classification task, standard binary class imbalance metrics do not apply. To provide a granular understanding of classification boundaries, Table 3 stratifies accuracy by syllable count. This breakdown effectively serves as the diagonal of confusion matrix, demonstrating model behavior across varying structural complexities.

The fusion of prosody and spectral features produces the most accurate results, both for Spanish (97.5%) and English (93.9%). In addition, the performance remains stable regardless of the number of syllables per word, highlighting the robustness of the trained models. These results establish an upper-bound baseline for model performance, since they are derived from in-domain, healthy, and native speech. The following subsection benchmarks the best-performing models (Prosody+Spectral) against out-of-domain datasets, specifically evaluating their generalization to pathological and non-native speech.

4.1 Benchmark Evaluation

The benchmark evaluation for the different out-of-domain test corpora is detailed in Table 4. The results include the overall accuracy of the stress detection mod-

Table 3. In-domain syllable stress detection accuracy (%) for Spanish (Common Voice) and English (LibriSpeech) models, comparing prosodic, spectral, and fused feature sets.

Feature set	Overall	# Syllables			
		2	3	4	5
Spanish - Common Voice					
Prosody	86.8	86.3	86.7	87.8	87.8
Spectral	90.3	88.9	89.8	92.7	95.2
Prosody+Spectral	97.5	91.1	92.6	94.7	96.7
English - Librispeech					
Prosody	88.3	90.5	84.7	77.0	81.7
Spectral	93.4	94.2	91.7	89.9	91.1
Prosody+Spectral	93.9	94.7	92.6	90.4	93.4

els across the full corpus (ALL), and for stratified groups within each dataset. Accuracies are further broken down by the number of syllables per word.

When evaluated on pathological and non-native speech benchmarks, the models demonstrate a reduction in performance compared to the in-domain healthy baseline (Table 3), reflecting the expected deviations in stress patterns. For instance, in the Spanish AphasiaBank partition, which exclusively contains patient data, the overall accuracy drops to 75.5%.

Regarding the English version of AphasiaBank, where there is information both from HC and patients, we observed a clear difference between the stress detection accuracies observed for HCs and patients. Moreover, such a difference becomes higher as long as the Aphasia severity increases. Note that patients with severe and very severe Aphasia did not produce words with 4 and 5 syllables. A more detailed evaluation performance for each subgroup of patients is shown in Fig. 2. These results confirm that an increase in disease severity leads not only to a reduction in the average number of correctly stressed syllables, but also to an increased variance among patients within the same aphasia severity.

On the Ivanova corpus, the model achieved an overall accuracy of 78.9%. While performance between HC and MCI patients remains comparable, accuracy for AD patients drops by roughly 2%. A similar trend is evident in the English ADReSS-M corpus, where AD patients scored slightly lower than HCs. We additionally evaluated whether stress detection probabilities correlated with the subjects' cognitive evaluations via their MMSE scores. However, no significant correlations were found in either corpus. Evaluation on the Neurovoz PD benchmark yielded an overall accuracy of 83.4%, with no significant differences observed between PD patients and HC subjects, and neither among patients in different H&Y stadiums. However, since this corpus contains varied speech tasks, an additional task-specific analysis was performed. Accuracy reached 85.5% during read sentences, but declines to 79.1% during spontaneous monologues.

Table 4. Cross-corpus evaluation of syllable stress detection accuracy (%). The performance is detailed for overall accuracy and by the number of syllables per word across healthy, pathological (Aphasia, AD, MCI, PD), and L2 learner populations.

		# Syllables			
Test set	Overall	2	3	4	5
AphasiaBank - Spanish					
Aphasia	75.5	74.0	79.7	76.6	100.0
AphasiaBank - English					
ALL	88.6	89.7	87.5	76.0	88.7
HC	90.5	91.9	88.2	79.5	87.3
Aphasia	86.4	87.1	86.5	68.8	92.1
Aphasia-Mild	87.0	87.8	87.3	71.4	92.0
Aphasia-Moderate	86.1	87.0	85.5	60.9	88.9
Aphasia-Severe	83.6	83.5	83.7	-	-
Aphasia-Very Severe	62.9	61.1	80.0	-	-
Ivanova - Spanish					
ALL	78.9	78.2	77.2	90.3	67.9
HC	79.2	78.8	77.2	90.6	67.4
MCI	79.4	78.1	78.0	91.0	75.4
AD	77.3	76.7	76.1	88.0	57.1
Neurovoz - Spanish					
ALL	83.4	85.0	77.7	83.2	80.4
HC	83.4	84.4	78.8	84.6	82.5
PD	83.3	85.4	76.1	80.0	66.7
ADReSS-M - English					
ALL	83.5	83.6	84.1	73.5	75.0
HC	84.2	84.2	85.1	74.4	75.0
AD	82.8	82.9	83.2	68.8	-
SPLOCC - Spanish					
ALL	71.8	69.8	75.8	76.4	70.1
Beginners (A2)	63.7	60.7	72.2	66.1	60.0
Intermediate (B1-B2)	68.4	66.2	72.8	73.3	66.0
Advanced (C1-C2)	73.3	71.2	76.7	78.0	71.1
Native	78.8	78.1	80.6	82.1	74.8

The observed disparity in model performance, where accuracy for PD patients remains comparable to HCs but declines for AD patients, warrants further discussion. PD is primarily a motor control disorder characterized by hypokinetic dysarthria. While these patients frequently exhibit a reduced prosodic range, such as monopitch and monoloudness, we hypothesize that their underlying cog-

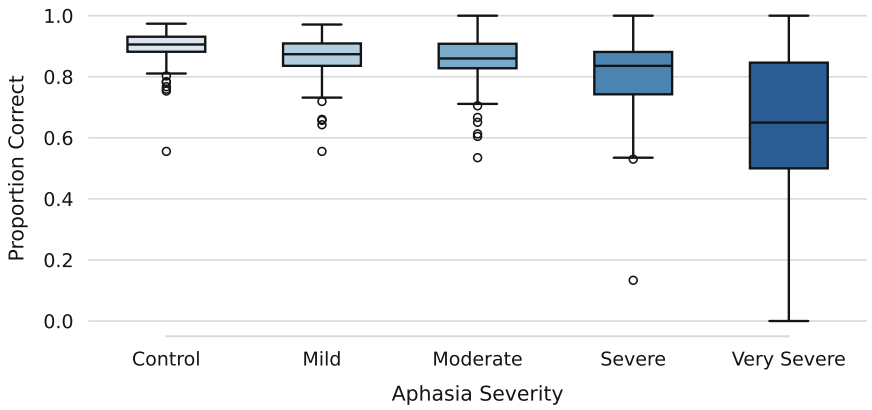

Fig. 2. Correctly stressed syllables by the Aphasia severity in the English AphasiaBank.

nitive intent remains intact, aligning with previous findings in [6]. Furthermore, studies indicate that speakers with motor speech disorders often apply compensatory acoustic strategies, such as increasing pitch, intensity, and duration to emphasize a word within a sentence [21]. Therefore, high detection accuracy is maintained for PD patients. Conversely, AD involve cortical neurodegeneration that directly affects cognitive and linguistic networks. The resulting semantic deficits, word-finding difficulties, and reading dysfluencies can disrupt the actual syllabic sequencing and rhythm of speech [5]. In these cognitive conditions, patients may place stress on the incorrect syllable or introduce atypical intra-word pauses, leading to deviations from canonical stress patterns [4]. This performance degradation may reflect underlying cognitive-linguistic disruptions rather than purely motor-based prosodic cues.

Finally, the evaluation of the L2 SPLOCC dataset showed a stress detection of 71.8%. In this case, a clear reduction in the stress detection is observed depending on the Spanish proficiency level of the students, degrading sequen-

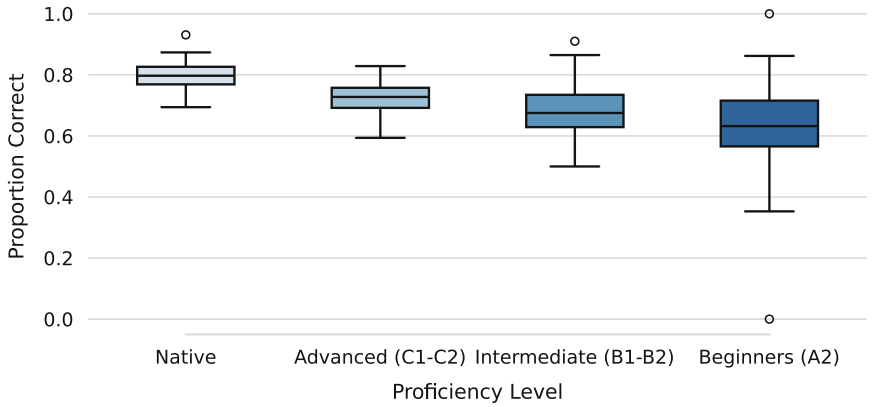

Fig. 3. Correctly stressed syllables by proficiency group in the SPLOCC corpus.

tially from native speakers (78.8%) down to A2-level beginners (63.7%). Figure 3 further illustrates the distribution of correctly stressed syllables across these proficiency groups. These results confirm the potential applicability of the proposed approach as a pronunciation level assistant in language learning systems.

5 Conclusions and Future Work

This paper introduced a pipeline for automated syllable stress detection in Spanish and English. By integrating 21 prosodic markers with 85 spectral and articulatory features, the proposed approach successfully captures the complex, suprasegmental nature of lexical stress. Our models achieved in-domain accuracies of 97.5% for Spanish and 93.9% for English, providing a strong baseline for evaluating non-canonical speech.

The benchmark evaluation revealed several key insights into the utility of automated stress detection for the assessment patients with neurological conditions. The model's performance remained stable for PD patients, likely due to their preserved cognitive intent regarding stress placement. On the contrary, a clear reduction in accuracy was observed as the severity of Aphasia increased. Similarly, the model captured distinct deviations in AD patients compared to HCs, suggesting that automated stress detection can reflect cognitive-linguistic disruptions. Regarding the L2 Proficiency Tracking, the evaluation on the SPLLOC corpus demonstrated that stress detection accuracy scales with the Spanish proficiency levels of L2 learners (A2 to native). These results validate the model's potential as an objective tool for providing localized feedback in language learning applications. For future work, we intend to implement this pipeline into a real-time feedback interface for clinical rehabilitation, allowing therapists to quantitatively track patient progress over longitudinal interventions.

Acknowledgments. This paper was funded by the Basque Government (Spri Group) through funding IKASPROD project (KK-2024/00050).

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

References

1. Aluru, S.H.: Post-net: a linguistically inspired sequence-dependent transformed neural architecture for automatic syllable stress detection. In: Interspeech, pp. 3340–3344 (2024)
2. Ardila, R.: Common voice: a massively-multilingual speech corpus. In: LREC, pp. pp. 4218–4222 (2020)
3. Bisani, M., Ney, H.: Joint-sequence models for grapheme-to-phoneme conversion. Speech Commun. **50**(5), 434–451 (2008)
4. del Carmen Perez-Sanchez, M., et al.: Reading fluency in Spanish patients with Alzheimer's disease. Curr. Alzheimer Res. **18**(3), 243–255 (2021)

5. Colombo, L., et al.: The impact of lexical-semantic impairment and of executive dysfunction on the word reading performance of patients with probable Alzheimer dementia. Neuropsychologia **42**(9), 1192–1202 (2004)
6. Exner, A.H., et al.: The effects of speech task on lexical stress in Parkinson's disease. Am. J. Speech Lang. Pathol. **32**(2), 506–522 (2023)
7. Goedemans, R.: The Study of Word Stress and Accent: Theories, Methods and Data. Cambridge University Press (2018)
8. Hernández-Figueroa, Z., et al.: Automatic syllabification for Spanish using lemmatization and derivation to solve the prefix's prominence issue. Expert Syst. Appl. **40**(17), 7122–7131 (2013)
9. Ivanova, O., et al.: Discriminating speech traits of Alzheimer's disease assessed through a corpus of reading task for Spanish language. CS L **73**, 101341 (2022)
10. Kim, J.W.: Crepe: a convolutional representation for pitch estimation. In: ICASSP, pp. 161–165. IEEE (2018)
11. Li, K., et al.: Automatic lexical stress and pitch accent detection for l2 english speech using multi-distribution deep neural networks. Speech Commun. **96**, 28–36 (2018)
12. Luz, S.: Multilingual Alzheimer's dementia recognition through spontaneous speech: a signal processing grand challenge. In: ICASSP, pp. 1–2 (2023)
13. MacWhinney, B., Fromm, D.: Aphasiabank as bigdata. In: Seminars in Speech and Language. vol. 37, pp. 010–022 (2016)
14. Mallela, J.: A comparative analysis of sequential models that integrate syllable dependency for automatic syllable stress detection. In: Interspeech (2024)
15. Mallela, J.: Uncovering lexical stress patterns: a layered analysis of self-supervised representations for automatic syllable stress detection. SLaTE., 177–181 (2025)
16. Mendes-Laureano, J., et al.: Neurovoz: a castillian Spanish corpus of parkinsonian speech. Scientific Data **11**(1), 1367 (2024)
17. Mendoza Ramos, V., et al.: Acoustic identification of sentence accent in speakers with dysarthria: cross-population validation and severity related patterns. Brain Sci. **11**(10), 1344 (2021)
18. Mitchell, R., et al.: Splloc: a new database for Spanish second language acquisition research. EUROSLA Yearb. **8**(1), 287–304 (2008)
19. Panayotov, V.: Librispeech: an ASR corpus based on public domain audio books. In: ICASSP, pp. 5206–5210 (2015)
20. Povey, D.: The kaldi speech recognition toolkit, ASRU (2011)
21. Tykalova, T., et al.: Acoustic investigation of stress patterns in Parkinson's disease. J. Voice **28**(1), e1–e129 (2014)
22. Yarra, C.: Automatic detection of syllable stress using sonority based prominence features for pronunciation evaluation. In: ICASSP, pp. 5845–5849 (2017)