Deep Learning-Based Emotion Detection in Aphasia Patients

David Ortiz-Perez[1], Pablo Ruiz-Ponce[1], Javier Rodríguez-Juan[1], David Tomás[1], Jose Garcia-Rodriguez[1](✉), and Grzegorz J. Nalepa[2]

[1] Universidad de Alicante, Alicante, Spain
{dortiz,pruiz,jrodriguez,jgarcia}@dtic.ua.es, dtd@ua.es
[2] Jagiellonian University, Kraków, Poland
grzegorz.j.nalepa@uj.edu.pl

Abstract. In this paper, we propose a pipeline for analyzing audio recordings of both aphasic and healthy patients. The pipeline can transcribe and distinguish between patients and the interviewer. To evaluate the pipeline's effectiveness, we conducted a manual review of the initial frames of one hundred randomly selected samples and achieved a 94% accuracy in patient differentiation. This evaluation aimed to ensure accurate differentiation when analyzing frames where the clinician interacts with the patient. This differentiation is important, as the primary objective of this project is to examine patients' emotions while they listen to their interviewer and identify patterns between healthy patients and those with aphasia. To achieve this, we used the AphasiaBank dataset, which includes video recordings of interviews with both aphasic and healthy patients. By combining the audio differentiation with the video recordings, we were able to analyze the facial expressions of patients while they listened to the speech of the interviewer. This analysis revealed a negative influence on the mood of aphasic patients. This negative influence stems from aphasic patients' difficulty in correctly understanding and expressing speech.

Keywords: Aphasia · Emotion recognition · Deep Learning · Transformers

1 Introduction

Aphasia is a neurological disorder that occurs due to damage to certain regions of the brain involved in speech and language. This condition can cause significant communication difficulties for patients, who may be unable to express themselves clearly. In the United States, it affects approximately one million people and is commonly associated with middle-aged and older individuals, though it can occur at any age.

Symptoms of aphasia primarily involve difficulties with language. There are various types of aphasia, which are determined by the specific location and extent of brain damage [10,13,15]. The most prevalent types are Wernicke, Broca, and

P. García Bringas et al. (Eds.): SOCO 2023, LNNS 750, pp. 195–204, 2023.
https://doi.org/10.1007/978-3-031-42536-3_19

global aphasia. Wernicke aphasia is characterized by the use of nonsensical, long sentences and the invention of new words, making it challenging for patients to comprehend others' speech. In contrast, Broca aphasia results in patients using minimal words and constructing short, direct sentences, frequently omitting common words such as "the", "and", or "is". Global aphasia involves extensive brain damage and is associated with severe communication difficulties that limit patients' ability to both speak and comprehend others' speech. Other less common types of aphasia affect patients' communication abilities differently. Aphasia can result from various conditions such as strokes, brain tumors, or progressive neurological diseases like Alzheimer's disease, which is often linked to dementia [5,7,18].

This study aims to develop a pipeline that can transcribe and distinguish between patient and clinician recordings for further analysis of patients' facial expressions while they listen to clinicians. The primary objective is to analyze patients' emotions to identify patterns in aphasia disease, particularly regarding how patients feel while listening to others, such as clinicians. Patients with aphasia may have different moods due to difficulty comprehending language. Analyzing their reactions and emotions can help improve communication with them, ultimately enhancing their comfort levels.

The remaining of the paper is organized as follows: Sect. 2 will introduce the state of the art and related work in this area; Sect. 3 will analyze the data available in our dataset; Sect. 4 will show the work done over the dataset for analyzing the results in Sect. 5; finally, in Sect. 6 we summarize our conclusions and propose further work.

2 Related Work

A research study has been conducted to select the most suitable dataset for our project. The only dataset that provides information regarding aphasia disease is AphasiaBank[1] [6], which will be explained in detail in Sect. 3 and is provided by TalkBank. TalkBank is primarily dedicated to the research of human communication and offers other similar datasets that have been considered for our project. One such dataset is TBIBank [4], which contains information on patients with traumatic brain injuries. This dataset is similar to the AphasiaBank corpus, as aphasia is often the result of brain damage in certain areas. Another comparable dataset is RHDBank [14], which contains information on patients with right hemisphere damage. Lastly, there is DementiaBank[11], which contains information on patients with dementia. Dementia is another cause of aphasia, as it is a progressive neurological disease.

The main factor that drove the selection of the AphasiaBank dataset over the others was the availability of video recordings and a larger number of samples. While DementiaBank only contains audio recordings and TBIBank do not provide a video modality for every sample, both AphasiaBank and RHDBank

[1] https://sla.talkbank.org/TBB/aphasia.

provide video recordings for each sample. Moreover, AphasiaBank offers a significantly larger number of samples for our study. In this research, the video modality is essential for emotion recognition, as it is easier to predict when the facial expressions of a person are visible.

Regarding the existing work carried out on the selected dataset, AphasiaBank, there are tasks such as automatic speech recognition of aphasic individuals, as well as numerous lexical and semantic analyses, as this dataset includes transcriptions of recordings.

The task of automatic speech recognition, which involves transcribing an audio recording, has shown significant advancements in recent years, particularly with transformer-based architectures such as Whisper [17] or Wav2Vec2 [1]. The significance of this area lies in the added complexity of the task due to the communication difficulties faced by aphasic patients who may produce incomprehensible speech or sentences during a conversation. Additionally, there is a significant disparity in the availability of transcription data for healthy patients compared to those with the disease. In this regard, we highly appreciate the work done by Iván G. Torres et al. [22], who used the AphasiaBank dataset as well.

Regarding other works focused on the semantic and lexical analysis of transcriptions from this dataset, several studies can be found. One such example is the work by Yu-Er Jiang et al. [9], which analyzed the main verbs and nouns used by patients with anomic aphasia and healthy controls. The study compared individuals of similar age and education levels to ensure a more accurate and balanced analysis. Results showed that individuals with anomic aphasia tend to use fewer core verbs and nouns than healthy individuals. Another study in this area that utilized the same dataset was conducted by Ouden Dirk-Bart et al. [16], which analyzed the use of verbs. Results showed that individuals with Broca's aphasia tend to use verbs in less complex and diverse ways than healthy individuals.

Emotional expressions and understanding are crucial in human communication. Diseases such as Aphasia and Dementia can negatively impact interactions and conversations with others. Patients with dementia may find it difficult to identify others' emotions and empathize with them. Thus, investigating the emotions of these patients is an interesting area of study. With advancements in artificial intelligence, tasks such as emotion recognition can be automated. Although there are currently no studies on how aphasia affects patients' emotions, diseases like dementia have been explored in automating emotion recognition for further analysis.

Karmele Lopez-de-Ipiña et al. [8] conducted an emotion response analysis aimed at detecting dementia by analyzing audio recordings and using audio features to determine emotions. Parkinson's disease is another illness that can affect patients' emotions, with deficits in emotional speech production. Shunan Zhao et al. [23] have performed a more complex analysis using automatic emotion recognition to investigate this disease.

In this context, there can be numerous emotions, with subtle differences between them. Psychologist Paul Ekman differentiates between six basic emotions: anger, disgust, happiness, fear, surprise, and sadness. Ekman proposed this distinction based on an analysis of eye, head, and facial muscle movements.

3 Dataset

As previously indicated, our dataset selection process culminated in the decision to use the AphasiaBank dataset. This particular dataset is of particular interest due to its inclusion of video recordings, where patients were recorded during a conversation with a clinician. The videos capture the upper half of the body, including the face, which makes the facial expressions of the patients the most crucial aspect for our analysis of emotional recognition.

The dataset includes video recordings of both healthy and aphasic patients, although there are considerably more samples from the latter group. The dataset contains a total of 440 video samples from aphasic patients and 220 samples from healthy patients. The primary focus of the recordings is on the speech behavior of the patient, with the conversation and discourse tasks designed to provide data on how they express themselves. Since this database has different corpuses, it is important to note that the tasks vary depending on the corpus of the dataset, and some tasks are more varied than others. A corpus is a set of data from the dataset, in this case, a set of video recordings. The main task involves initiating a conversation by inquiring about the patient's perception of their speech, while other tasks include the description of various images.

The dataset also includes CHAT transcriptions [12] of the conversations, which is in line with other similar datasets, such as DementiaBank. In this sense, the information represented in the form of text that captures the speech that has been performed can be highly valuable for semantic and lexical analysis, as demonstrated in previous studies.

4 Approach

In this project, we developed a pipeline for automatic speech recognition and speaker differentiation of the video recordings in the AphasiaBank dataset. The pipeline consists of several stages and has been applied to each sample of the dataset. First, we extract the audio information from the video and store it. Using the Whisper model developed by OpenAI, we transcribe the recording, resulting in two files: a plain transcription file and a file with transcription and time-lapse of the transcripted sentences. The latter is used for further processing.

Next, we use the speaker-diarization [2,3] model provided by the HuggingFace library to differentiate between the patient and the clinician. This model enables us to obtain a time-lapse of when each speaker is talking. Both models are transformer-based, which has significantly improved the accuracy of the pipeline. Using the output from both models, we obtain a final transcription of what each speaker says. In order to distinguish between the patient and the clinician, we propose to identify the patient as the person who speaks for a longer duration in the recordings. This approach is based on the fact that the recordings are primarily focused on the speech of the patients, who are expected to speak more than the clinicians. In this scenario, the role of the clinicians is to facilitate the conversation and provide assistance to the patients when necessary.

For emotion recognition, we extract the time-lapse where the patient is listening to the clinician, and only keep the video frames during this period. The pipeline architecture is shown in the Fig. 1. Overall, our pipeline provides an efficient and accurate method for processing audio recordings and extracting important information for further analysis.

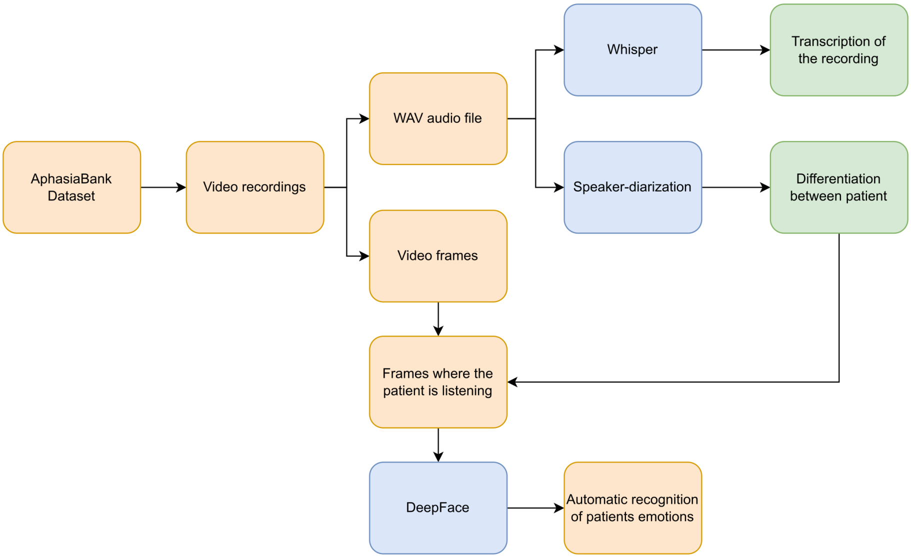
AphasiaBank Dataset
Video recordings
WAV audio file
Video frames
Whisper
Speaker-diarization
Transcription of the recording
Differentiation between patient
Frames where the patient is listening
DeepFace
Automatic recognition of patients emotions

Fig. 1. Architecture of the proposed pipeline

Once we have identified the video frames where the patient is listening to the clinician, we utilize the DeepFace [19–21] library's model to extract relevant information from facial expressions. While this model can provide information about age, sex, and race, our focus is solely on the emotions conveyed through facial expressions. The model identifies emotions such as anger, disgust, fear, happiness, sadness, surprise, and neutral. Those emotions are the previously mentioned in Sect. 2. We will use this information to develop a method for analyzing the emotions conveyed in each sample. The other relevant information obtained through transcription and speaker differentiation with time lapses will not be used in this project. However, we will keep this information for future works.

5 Evaluation

In order to evaluate this work, the diarisation component was tested as the first step. One hundred random samples were selected from the dataset, with patients and clinicians properly differentiated, and were manually reviewed. The pipeline was tested by analyzing the initial few minutes of the selected samples along with the diarisation, as the entire files were not analyzed due to some samples being up to an hour long. The pipeline correctly distinguished ninety-four out of the one hundred samples. However, the remaining six samples were incorrectly labeled in terms of differentiating between the clinician and the patient while they were speaking. As a result, the accuracy rate in distinguishing between patients and clinicians was 94%. This differentiation task is important for the ultimate goal of analyzing the facial expressions of patients while they are listening to the clinician. The samples that were incorrectly distinguished were those where the clinician had to speak extensively to maintain the conversation and assist the patients who were not able to communicate fluently. In such cases, the pipeline identified the clinician as a patient since it considers the person who speaks more as the patient. Additionally, the low recording quality was another reason for incorrect labeling. Nonetheless, the pipeline generally distinguishes the majority of cases correctly. On the other hand, no evaluation has been done over the transcription text, since it has not finally used in this work.

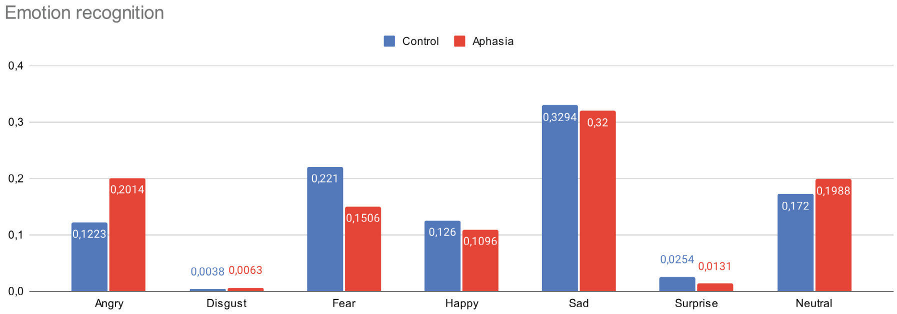

Fig. 2. Mean of emotions represented in the analysis over patients while listening to clinicians' speech

The other evaluation metric involved comparing the results obtained from both aphasic and healthy patients in terms of emotion recognition. The results are shown in the Fig. 2 and in more detail in Table 1, with an individual percentage for each corpus provided in the dataset. These metrics represent the average emotions displayed by the patients' facial expressions during the interview, represented as a percentage between zero and one, for example, 0.5 represents the half of every represented emotion. These metrics are The most notable difference was observed in the mean value of the "angriness" emotion. This finding was not surprising, as patients may experience frustration and anger due to difficulties in understanding the speech of the clinician. Similarly, although it represents a small proportion of the mean of the emotions, the aphasic patients showed double the proportion of "disgust" emotion compared to the healthy patients. Other significant differences were observed in the proportions of "fear", "surprise", and "neutrality" emotions. The lower proportion of "fear" and "surprise" and the higher proportion of "neutral" emotion may be due to the difficulty in understanding the speech. In the case of not understanding the clinician's speech, patients may not show fear or surprise as healthy patients would when they fully comprehend a sentence and are surprised by its content. Additionally, the higher proportion of "neutral" emotion may result from the lack of expression due to poor speech recognition.

Table 1. Average emotion detection in the different corpuses of the dataset

Corpus		Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral
Control	Wright	0.153	0.021	0.238	0.041	0.409	0.044	0.094
	Capilouto	0.053	0.003	0.121	0.001	0.816	0.000	0.006
	Kempler	0.019	0.005	0.262	0.224	0.459	0.008	0.022
	Richardson	0.000	0.000	0.062	0.655	0.130	0.000	0.153
	MSU	0.083	0.000	0.088	0.139	0.475	0.001	0.214
	Total	**0.122**	**0.003**	**0.221**	**0.126**	**0.329**	**0.025**	**0.172**
Aphasia	Wright	0.160	0.000	0.137	0.343	0.127	0.001	0.232
	Thompson	0.113	0.000	0.105	0.347	0.169	0.040	0.226
	Adler	0.088	0.000	0.532	0.041	0.105	0.046	0.187
	UNH	0.344	0.001	0.343	0.045	0.235	0.001	0.031
	STAR	0.554	0.008	0.074	0.021	0.333	0.000	0.011
	TAP	0.359	0.019	0.267	0.032	0.265	0.002	0.056
	Garrett	0.051	0.000	0.220	0.212	0.068	0.000	0.449
	Whiteside	0.350	0.001	0.153	0.143	0.221	0.031	0.100
	Tucson	0.114	0.000	0.066	0.101	0.481	0.001	0.238
	Fridriksson	0.040	0.000	0.109	0.050	0.450	0.007	0.343
	UCL	0.296	0.000	0.139	0.024	0.198	0.029	0.313
	TCU	0.137	0.001	0.073	0.025	0.742	0.000	0.022
	Elman	0.256	0.000	0.089	0.085	0.549	0.000	0.021
	CMU	0.231	0.000	0.132	0.477	0.059	0.002	0.098
	Kurland	0.572	0.001	0.035	0.092	0.250	0.000	0.050
	TCU-bi	0.079	0.000	0.062	0.339	0.316	0.000	0.204
	Kempler	0.189	0.006	0.067	0.388	0.298	0.004	0.048
	Kansas	0.132	0.000	0.254	0.113	0.373	0.000	0.127
	SCALE	0.296	0.004	0.109	0.112	0.261	0.016	0.201
	ACWT	0.119	0.004	0.072	0.341	0.100	0.017	0.347
	Wozniak	0.266	0.002	0.088	0.039	0.322	0.041	0.242
	MSU	0.064	0.004	0.014	0.073	0.484	0.000	0.361
	Williamson	0.025	0.001	0.001	0.019	0.226	0.000	0.728
	BU	0.509	0.002	0.152	0.053	0.127	0.024	0.134
	Total	**0.201**	**0.006**	**0.150**	**0.109**	**0.32**	**0.013**	**0.198**

6 Conclusion

This work proposes a pipeline to analyze video recordings of Aphasia patients and obtain time-lapses of moments where patients are listening to their interviewer. The pipeline aims to obtain time-lapses of moments where patients are listening to their interviewer. To achieve this goal, the study conducted research in the area of Automatic Speech Recognition tasks and differentiation between speakers. Based on this research, the study selected two models, namely Whisper and speaker-diarization, to develop the pipeline.

The effectiveness of the pipeline was evaluated by manually reviewing the beginning of one hundred randomly selected video samples from the dataset used. The pipeline was also used to recognize emotions in both healthy and Aphasia patients. The DeepFace library was utilized to detect emotions from the facial expressions of patients. The study found that Aphasia patients express different emotions than healthy patients when listening to someone's speech, mainly due to their difficulties in understanding and expressing speech, which negatively influences their mood. This analysis of their emotional state can help improve their interactions by avoiding conversations that may have a negative impact on their mood.

Future work in this field includes proposing and deploying a more complex system for analyzing patients' facial expressions. The new system would include additional features beyond emotions to identify other facial expression patterns between healthy and Aphasia patients. Another idea is to analyze the transcription of the different samples to identify patterns in what they express and are listening that could lead to further interesting research. With this transcription analysis, a deeper emotional analysis can be implemented to identify the type of speech that has a negative impact on their mood. Finally, the study plans to expand the project to include other similar diseases, such as Traumatic Brain Injuries, to explore their effects on patients.

Acknowledgment. We would like to thank "A way of making Europe" European Regional Development Fund (ERDF) and MCIN/AEI/10.13039/501100011033 for supporting this work under the TED2021-130890B (CHAN-TWIN) research Project funded by MCIN/AEI /10.13039/501100011033 and European Union NextGenerationEU/PRTR, MoDeaAS project (grant PID2019-104818RB-I00) and AICARE project (grant SPID202200X139779IV0). Also the HORIZON-MSCA-2021-SE-0 action number: 101086387, REMARKABLE, Rural Environmental Monitoring via ultra wide-ARea networKs And distriButed federated Learning. Furthermore, we would like to thank Nvidia for their generous hardware donation that made these experiments possible.

References

1. Baevski, A., Zhou, H., Mohamed, A., Auli, M.: wav2vec 2.0: a framework for self-supervised learning of speech representations (2020)
2. Bredin, H., Laurent, A.: End-to-end speaker segmentation for overlap-aware resegmentation. In: Proceedings of Interspeech 2021, Brno, Czech Republic (2021)

3. Bredin, H., et al.: Pyannote. audio: neural building blocks for speaker diarization. In: ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing, Barcelona, Spain (2020)
4. Elbourn, E., Kenny, B., Power, E., Togher, L.: Psychosocial outcomes of severe traumatic brain injury in relation to discourse recovery: a longitudinal study up to 1 year post-injury. Am. J. Speech-Lang. Pathol. **28**, 1–16 (2019). https://doi.org/10.1044/2019_AJSLP-18-0204
5. Fernández Montenegro, J.M., Villarini, B., Angelopoulou, A., Kapetanios, E., Garcia-Rodriguez, J., Argyriou, V.: A survey of Alzheimer's disease early diagnosis methods for cognitive assessment. Sensors **20**(24) (2020). https://doi.org/10.3390/s20247292. https://www.mdpi.com/1424-8220/20/24/7292
6. Forbes, M., Fromm, D., Macwhinney, B.: Aphasiabank: a resource for clinicians. In: Seminars in Speech and Language, vol. 33, pp. 217–22 (2012). https://doi.org/10.1055/s-0032-1320041
7. Gomez-Donoso, F., et al.: A robotic platform for customized and interactive rehabilitation of persons with disabilities. Pattern Recognit. Lett. **99**, 105–113 (2017). https://doi.org/10.1016/j.patrec.2017.05.027. https://www.sciencedirect.com/science/article/pii/S0167865517301903. User Profiling and Behavior Adaptation for Human-Robot Interaction
8. López-de Ipiña, K., et al.: On the selection of non-invasive methods based on speech analysis oriented to automatic Alzheimer disease diagnosis. Sensors **13**(5), 6730–6745 (2013). https://doi.org/10.3390/s130506730. https://www.mdpi.com/1424-8220/13/5/6730
9. Jiang, Y.E., Liao, X.Y., Liu, N.: Applying core lexicon analysis in patients with anomic aphasia: based on mandarin aphasiabank. Int. J. Lang. Commun. Disord. (2023). https://doi.org/10.1111/1460-6984.12864. https://onlinelibrary.wiley.com/doi/abs/10.1111/1460-6984.12864
10. Johns Hopkins Medicine: Aphasia. https://www.hopkinsmedicine.org/health/conditions-and-diseases/aphasia
11. Lanzi, A., Saylor, A., Fromm, D., Liu, H., Macwhinney, B., Cohen, M.: Dementiabank: theoretical rationale, protocol, and illustrative analyses. Am. J. Speech-Lang. Pathol. **32**, 1–13 (2023). https://doi.org/10.1044/2022_AJSLP-22-00281
12. Macwhinney, B.: The childes project: tools for analyzing talk. Child Lang. Teach. Ther. **8** (2000). https://doi.org/10.1177/026565909200800211
13. Mayo Clinic: Aphasia (2022). https://www.mayoclinic.org/diseases-conditions/aphasia/symptoms-causes/syc-20369518
14. Minga, J., Johnson, M., Blake, M., Fromm, D., Macwhinney, B.: Making sense of right hemisphere discourse using RHDBank. Top. Lang. Disord. **41**, 99–122 (2021). https://doi.org/10.1097/TLD.0000000000000244
15. National Institute of Mental Health: What is aphasia? - types, causes and treatment. https://www.nidcd.nih.gov/health/aphasia
16. Ouden, D.B., Malyutina, S., Richardson, J.: Verb argument structure in narrative speech: mining the AphasiaBank. Front. Psychol. **6** (2015). https://doi.org/10.3389/conf.fpsyg.2015.65.00085
17. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision (2022)

18. Revuelta, F.F., Chamizo, J.M.G., Garcia-Rodrguez, J., Sáez, A.H.: Representation of 2D objects with a topology preserving network. In: Quereda, J.M.I., Micó, L. (eds.) Pattern Recognition in Information Systems, Proceedings of the 2nd International Workshop on Pattern Recognition in Information Systems, PRIS 2002, In conjunction with ICEIS 2002, Ciudad Real, Spain, April 2002, pp. 267–276. ICEIS Press (2002)
19. Serengil, S.I., Ozpinar, A.: Lightface: a hybrid deep face recognition framework. In: 2020 Innovations in Intelligent Systems and Applications Conference (ASYU), pp. 23–27. IEEE (2020). https://doi.org/10.1109/ASYU50717.2020.9259802
20. Serengil, S.I., Ozpinar, A.: Hyperextended lightface: a facial attribute analysis framework. In: 2021 International Conference on Engineering and Emerging Technologies (ICEET), pp. 1–4. IEEE (2021). https://doi.org/10.1109/ICEET53442.2021.9659697
21. Serengil, S.I., Ozpinar, A.: An evaluation of SQL and NOSQL databases for facial recognition pipelines (2023). https://www.cambridge.org/engage/coe/article-details/63f3e5541d2d184063d4f569. https://doi.org/10.33774/coe-2023-18rcn
22. Torre, I.G., Romero, M., Álvarez, A.: Improving aphasic speech recognition by using novel semi-supervised learning methods on aphasiabank for English and Spanish. Appl. Sci. **11**(19) (2021). https://doi.org/10.3390/app11198872. https://www.mdpi.com/2076-3417/11/19/8872
23. Zhao, S., Rudzicz, F., Carvalho, L.G., Marquez-Chin, C., Livingstone, S.: Automatic detection of expressed emotion in Parkinson's disease. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4813–4817 (2014). https://doi.org/10.1109/ICASSP.2014.6854516