INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Cross-Lingual Features for Alzheimer's Dementia Detection from Speech

Thomas Melistas[1], *Lefteris Kapelonis*[1], *Nikos Antoniou*[1], *Petros Mitseas*[1], *Dimitris Sgouropoulos*[1]
Theodoros Giannakopoulos[1,3], *Athanasios Katsamanis*[1,2], *Shrikanth Narayanan*[1,4]

[1] Behavioral Signal Technologies Inc., Los Angeles, CA, USA
[2] Institute for Language and Speech Processing, Athena Research Center, Athens, Greece
[3] NCSR Demokritos, Athens, Greece
[4] SAIL-University of Southern California, Los Angeles, CA, USA

{thomas, lefteris, nikos.antoniou, petros, dimitris, thodoris, nassos, shri}@behavioralsignals.com

Abstract

Alzheimer's dementia is a neurodegenerative disease that affects millions of people worldwide. Early detection of Alzheimer's dementia is crucial for effective treatment and management of the disease. In this paper, we present a cross-lingual approach for detecting Alzheimer's dementia from speech, based on multiple feature streams that capture the individual's speech and conversational interactions. In order to validate the ability of the features to perform well in cross-linguistic scenarios, we evaluate in a zero-shot setup, where the target domain is a language that was not available during training and a few-shot setup, where only limited data is available. Experimental results show that an ensemble system using the features trained on English and evaluated on Greek outperforms the baseline system by 4.4 %. Further experiments show promising zero-shot and few-shot performance on a similar Spanish task.
Index Terms: Alzheimer's Dementia, language-independent features, cross-lingual speech recognition

1. Introduction

Dementia is a disorder characterized by a long-term decrease of a wide range of cognitive functionalities: progressive impairments in memory, thinking and behaviors. Alzheimer's disease (AD) is the most common cause of dementia and can be challenging to diagnose [1]. While neuropsychological tests such as the mini-mental state examination (MMSE) [2] are commonly used for diagnosis, their results may not always be reliable [3]. To overcome this challenge, researchers have turned to signal processing and machine learning (ML) techniques to develop accurate and efficient methods for AD detection [4].

An essential condition to employ automated detection techniques is the collection of suitable speech corpora which can be utilized in order to train and evaluate ML models. The Pitt corpus from the Dementiabank [1] is a widely used publicly available speech dataset [5]. To address biases and imbalances in the data, the ADReSS Challenge [6] recently introduced a standardized version of the dataset. Furthermore, the ADReSSo Challenge [7] mandated solving the problem using only speech data without manual text transcriptions.

To address the challenges, models utilize features from both audio and text modalities, which are derived from either domain knowledge (disfluencies frequently exhibited by AD patients) or pre-trained models [8, 9, 10, 11, 12, 13]. A comparative study demonstrated that fine-tuning text transformers such as BERT can achieve higher accuracy than using handcrafted features [14]. In a recent work [15], an ensemble of different models was employed with speech paralinguistic, deep acoustic and text features. Another study [10] explored cognitive features that characterize losses of train of thought in addition to linguistic, speech, and disfluency features such as pauses and word rate. In their work, Yuan et al. [16] achieved high accuracy by utilizing the transformer model ERNIE along with encoding of pauses.

All the previously mentioned works focus on detecting AD using English language speech samples, training and evaluating ML models on English speaking subjects. The prevalence of English language in related works leads to some concern in whether the proposed methods can be transferred in other languages, possibly with quite different characteristics. While a few other studies, such as those using Mandarin Chinese [17] and Spanish [18], have explored the use of other languages, they still take a monolingual perspective, making their approaches language-specific, tailored exactly for speech samples from languages that they were trained on. However, a more demanding task would be to find powerful and predictive features that transfer across languages. Gosztolya et al. [19] extracted hesitation speech markers for both English and Hungarian, training separate ASR systems for each language. Interestingly, they found that even if the ASR systems were interchanged, the predictive power of these markers remained high. In contrast, Pérez-Toro et al. [20] take a different approach to study cross-lingual capabilities. They pre-train a model for AD detection in English, which is then transferred to achieve higher performance in Spanish-speaking subjects.

With regards to the issue of cross-lingual performance robustness, for example methods [19, 20] assumed that, during training, samples from both languages would be available. In this work, we depart from this assumption to study the cross-lingual problem on more challenging scenarios, to explore the cross-lingual capabilities of certain features. In the first setting, we assume that the models are trained on a source language (English) and evaluated on a different language (target), in a zero-shot manner. Our work for this setting started during our participation in the ICASSP 2023 Signal Processing Grand Challenge "ADReSS-M" [21], where the organizers provided a dataset with recordings of English-speaking subjects, whereas the validation and test sets contain samples from Greek participants. Extending the work conducted in this challenge, we investigate the performance of our method in the Spanish language in both zero-shot and few-shot manners, i.e. when no instances at all, or just a few instances are available from the

[1] https://dementia.talkbank.org/

10.21437/Interspeech.2023-1934

target language.

In both of these experimental settings, our proposed method combines four basic sets of signals, namely: stop-word ratio based on speech-to-text, paralinguistic information (modelled by low-level features and high-level outputs of models such as emotion, arousal and speaking rate), speaker interaction dynamics (utilizing speaker diarization predictions), along with demographic information about the patients.

In summary, our contributions are: 1) language-independent features that do not require manual transcripts, 2) an ensemble system that, using the above features, achieves an absolute improvement of 4.4% compared to the baseline on the "ADReSS-M challenge" and 3) zero-shot and few-shot evaluation of the features on a different-domain Spanish dataset. [2]

2. Methodology

2.1. Datasets

In our work, we used three distinct sources of data, containing speech from different languages, namely English, Greek and Spanish. In this Section, we provide more details regarding the datasets used in our study. In Table 1, we illustrate demographic information for the subjects that participated in the collection of each dataset.

2.1.1. English

For English, we used the training data provided by the organizers of [21]. This set contains descriptions of the Cookie Theft picture [5], with some recordings originating from Dementiabank. The dataset contains a total of 237 recordings, where 122 of them (51.47%) are labelled as "Probable AD", with the rest (115) corresponding to healthy subjects. An important challenge associated with these recordings is that they contain speech content from both the patient and the interviewer. We treat the audios in this form, without applying any manual preprocessing steps. Their mean duration is 76.9 seconds, with a standard deviation of 37.4 seconds.

2.1.2. Greek

For Greek, the overall setting is quite similar to the English dataset, the Greek-speaking subjects are describing a picture which represents a lion lying with a cub in the dessert while eating. Likewise, this dataset was obtained from the ADReSS-M challenge. It contains two subsets, one for validation purposes (8 samples) and one for testing (46 samples). The validation set contains 4 AD and 4 Healthy Control (HC) samples but the labels from the test set remained undisclosed. On average, the files from this dataset have a duration of 38 ± 20.1 seconds.

2.1.3. Spanish

In the case of Spanish, we utilize the dataset assembled by Ivanova et al. [22], also present in Dementiabank. This set contains recordings from subjects performing a standardized reading task. The subjects were asked to read a paragraph from a Spanish novel. This collection of data is consisted of 197 samples labelled as Healthy Control (HC), and 74 samples labelled as AD. However, there are 91 audio files coming from a third category, which characterizes patients with Mild Cognitive Impairment (MCI). We dismiss this class in order conform with the English and Greek datasets which contain only the two previous labels. The average duration of the samples is 46.63 seconds (with a standard deviation of 24 seconds).

One interesting aspect for the datasets chosen in our study is their differences in terms of spontaneity. In particular, the English and Greek datasets contain purely spontaneous speech, since participants are asked to describe a picture without any prior preparation. In the Spanish dataset, however, the degree of speech spontaneity and naturalness is limited since the subjects read a paragraph verbatim. From a ML viewpoint, this discrepancy poses another great challenge, alongside the need of good performance across languages, which makes the task of finding generalizable features even harder.

2.2. Feature Extraction

In this section, we describe the feature extraction procedure. For this purpose, we have adopted 6 feature sets from 4 different modalities.

2.2.1. Interaction dynamics features (Utt)

Timestamps for the speakers' utterances were extracted using the publicly available diarization system of PyAnnote Audio library [23]. The model's speaker timestamps provide valuable information, and they are extracted in a fully automated fashion. In other studies, this segmentation step was performed manually according to the timestamps provided in the dataset, however this is impractical and as the dataset size grows we expect this information to be absent. We extract a total of 7 features: 5 of them are related to utterance duration statistics (min, max, median, mean, std). Additionally, we have included one feature for the ratio of speech duration to the entire audio duration, as well as one for the average duration of pauses between utterances.

2.2.2. Paralinguistics: Hand-crafted features (Au:H)

Low level, hand-crafted audio features were extracted through the use of the pyAudioAnalysis library [24]. In particular, we have used 64 frame level features, including MFCCs, ZCR, Spectral Centroid and Chroma features and corresponding segment-level statistics (mean and std). We have adopted a short-term window size of 50msecs and a segment size of 1 second. These acoustic features are computed per utterance and then aggregated per file through a weighted sum, where each feature is weighted based on the duration of its utterance.

2.2.3. Paralinguistics: TRILLsson embeddings (Au:Tr)

We use the TRILLsson model [25] to get a 1024-dimension embedding for each speech segment using the previously mentioned speaker utterances (see Interaction Dynamics) and then compute the weighted average for the whole recording, using weights proportional to the respective segment durations. Specifically we use the second largest publicly available model (the 4th variant) which utilizes the Audio Spectrogram Transformer (AST) [26] architecture, since it performs best on the tasks of dysarthria detection and speech emotion recognition.

2.2.4. Speaking rate & arousal embeddings (Au:Sr)

Speech-behavioral features were computed from pre-trained CNN-based classifiers that recognize speaking rate (slow, normal, fast) and speech arousal (weak, normal, strong). The clas-

[2]The code is available here: `https://bitbucket.org/behavioralsignals/adress-m-2023/`

Lang.		Count	Years	Education Years	MMSE
ENG	AD Label - F/M	79/43	69.92 (6.36) / 68.37 (7.59)	11.51 (2.37) / 12.79 (2.85)	17.36 (5.07) / 18.72 (6.01)
	HC Label - F/M	75/40	65.62 (6.18) / 66.86 (6.40)	14.02 (2.67) / 13.92 (2.56)	29.00 (1.29) / 28.90 (0.9)
GR (val.)	AD Label - F/M	3/1	78.67 (1.70) / 69 (-)	7.67 (7.04) / 16 (-)	23.67 (2.87) / 29 (-)
	HC Label - F/M	3/1	66.33 (6.34) / 69 (-)	8.33 (3.30) / 16 (-)	29.33 (0.47) / 26 (-)
GR (test.)	F/M (No labels)	35/11	69.43 (7.28) / 67.55 (9.13)	11.57 (3.90) / 10.73 (3.86)	-
SP	AD Label - F/M	44/30	79.70 (7.98) / 79.17 (7.69)	7.78 (3.05) / 10.27 (4.80)	18.91 (4.66) / 21.53 (5.40)
	HC Label - F/M	139/58	75.62 (7.39) / 75.22 (9.03)	9.41 (3.77) / 9.97 (3.96)	28.18 (1.97) / 28.43 (1.64)

Table 1: *Mean (std) values for metadata information. F (M) stands for Female (Male), AD for Alzheimer's Dementia, HC for Healthy Control. We report these values per AD label (AD or HC), except for the Greek test set, for which the label information isn't available.*

Attempt	Model	ASR	Utt	Au:H	Au:Tr	Au:Sr	MD	F1-CV (%)	Acc-Val (%)	Acc-Test (%)
Attempt 1	SVM	✓				✓	✓	72.1	75.0	63.0
Attempt 2	SVM	✓	✓				✓	77.6	87.5	63.0
Attempt 3	XGB	✓		✓	✓	✓	✓	70.5	50.0	69.6
Attempt 4	XGB	✓	✓			✓	✓	74.3	75.0	71.7
Attempt 5	Majority Vote							-	75.0	78.3

Table 2: *Cross-Validation F1 on English train set (F1-CV), Accuracy on Greek validation and test sets.*

sifiers act on fixed-sized segments (3-sec) of the input audio. Similarly to previous sets of features, we obtain the per-class posterior (3 per task) for all 3-sec segments of the calculated utterances, and then we aggregate them by calculating their mean and std, amounting to 6 features per task (12 in total).

2.2.5. Stop-word ratio (ASR)

We obtained transcriptions using OpenAI's Whisper [27] medium model, which demonstrated adequate performance in transcribing the train set. Whisper has support for both Spanish and Greek languages, hence it can be used for multilingual transcription with relatively low WER. After obtaining the transcripts, we extract the percentage of stop-words as a feature, where we utilized the lexicons of SpaCy library [28] as collections of common stop words for English, Greek and Spanish languages. Similarly to the segmentation step, it is important to note that our method does not require any ground truth transcripts for each recording. The automated extraction of transcriptions may introduce some noise due to WER, but it is far more realistic as dataset size grows and only speech data are available.

2.2.6. Metadata (MD)

For this set of features, we included the age, gender and educational level of the patients as found in the metadata files of each audio recording. Missing educational levels were replaced by the respective mean value calculated in the training dataset set.

3. Experimental Results

In this section, we discuss the experiments that were conducted in order to evaluate the performance of our proposed features, in the task of AD classification.

3.1. Zero-Shot Evaluation on Greek

In this setup, the models were trained on the English dataset from "ADReSS-M" [21] and evaluated on the Greek validation and test sets. This evaluation is conducted in a zero-shot manner, since no data from the target language are seen during training. While it would be valuable to evaluate the proposed features in a few-shot manner, we were unable to do so due to the absence of ground truth labels for the Greek test set. We mainly use test set accuracy as the metric of comparison. At the time of development, the groundtruth labels from test set were unknown. To benchmark performance, we consider as baseline the model proposed by the challenge's organizers. The baseline approach achieved 73.9% classification accuracy, utilizing the eGeMAPS [29] speech features and the metadata information.

Our approach was to train numerous models by selectively concatenating features from the different modalities described in Section 2. We also experimented with applying PCA in order reduce the feature dimensionality and standardization to zero mean and unit variance. In this specific setup, since our goal is to maximize performance on the challenge's test set, we utilized a grid search strategy over the following hyperparameters to achieve maximum performance per individual model:

- Modalities of input features to include per model
- No PCA, PCA (10, 20, 30 components)
- Classifier: Support Vector Machines (SVMs) or XGBoost
- Feature scaling based on train, validation or test set statistics

Our final selection included 4 different models based on the cross-validation performance on the train set and accuracy on the validation set. We made sure to include each modality in at least one model to encapsulate as much information as possible and exploit the inherent value each had to offer. The final model was a voting ensemble of the four selected models. Its prediction corresponded to the majority vote and ties were resolved by selecting the class with the highest sum of posterior probabilities across all 4 models of the ensemble.

The results of these experiments are displayed in Table 2. Our best model, which is the ensemble of the first four models, achieves 78.3% classification accuracy, outperforming the baseline method by an absolute of 4.4%. The test set performance of our method was disclosed to us by the challenge organizers.

3.2. Evaluation on Spanish

To further evaluate the cross-lingual generalization ability of our features, we conducted zero-shot and few-shot evaluation

Features	CV F1 (%)	Zero-shot F1 (%)	Few-shot F1 (%) (10%)	Few-shot F1 (%) (20%)	Few-shot F1 (%) (50%)
Au:H	57.4	56.6	58.5	61.4	65.2
Au:H,Utt	63.6	58.2	60.1	62.5	**67.0**
Au:H,Utt,Au:Sr	64.2	59.6	61.2	63.5	66.1
Au:H,Utt,Au:Sr, ASR	64.7	60.4	62.2	64.3	65.2
Au:H,Utt,Au:Sr, ASR,MD	64.8	**62.5**	63.7	65.0	66.4
Au:H,Utt,Au:Sr, ASR,MD,Au:Tr	**67.0**	59.7	**64.2**	**66.2**	66.0

Table 3: *Cross-Validation (CV) F1, Zero-shot and Few-shot F1 evaluation on the Spanish dataset.*

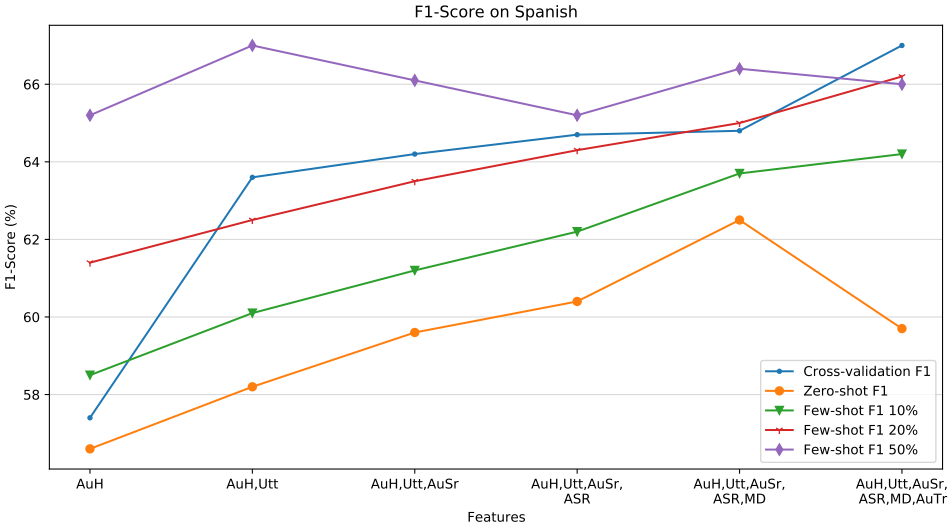

Figure 1: *F1-Score on Spanish language for various methods of training (cross-val on Spanish, Zero-shot, Few-shot for varying rates of available Spanish Data). x-axis denotes the available features (added incrementally) for each trial.*

experiments on the Spanish dataset. For the zero-shot evaluation, we trained SVMs on the English dataset and evaluated on the Spanish dataset. For the few-shot experiment, we partition the Spanish dataset into equal-sized folds. We then incorporate each fold in the training data along with English and evaluate on the remaining folds. The reported performance is averaged across the folds. To understand the impact of sample-efficiency, we conducted few-shot evaluations using varying sizes of available Spanish data. The macro-averaged F1 score was used to measure performance for all evaluations.

In Table 3, we present the results of our experiments. Despite the low performance achieved with cross-validation on the Spanish dataset, which may be due to the limited number of samples, our features demonstrated good cross-lingual generalization ability. The zero-shot evaluation showed only a small drop in performance compared to the cross-validation performance. When all features except Au:Tr are utilized, the F1 score in the zero-shot scenario is only 4.5% lower than the F1 score obtained through cross-validation. Additionally, incorporating only 20% of the Spanish samples and all features results in a decrease in F1 score by just 0.8%.

4. Conclusions

In this work, we present a cross-lingual approach for detecting Alzheimer's dementia from speech. Our approach is evaluated in diverse scenarios, where speech is both spontaneous (for Greek, English) and scripted (for Spanish). We propose extracting and fusing multiple feature streams that capture the individual's speech and conversational interactions. We evaluate our proposed method in zero-shot and few-shot experimental scenarios, to validate its ability to perform well in cross-linguistic setups. In the Greek dataset, our proposed set of features achieves 78.3% zero-shot classification accuracy, leading to an absolute increase of 4.4% when compared to the baseline approach from the organizers of ADReSS-M challenge [21]. For Spanish, experimental results prove that the proposed features lead to just a small drop of the performance in the zero shot scenario despite the difference in the task (lack of speech spontaneity). When only 20% of the target domain are used to tune the models, the achieved performance is almost equivalent to the cross validation performance in the target domain, which proves the generalization ability of the features. In the future, we plan to work on adding crosslingual semantics-related features and further explore which features transfer across tasks (spontaneous vs scripted). It would also be interesting to explore our proposed features on data from other languages.

5. References

[1] Z. Breijyeh and R. Karaman, "Comprehensive Review on Alzheimer's Disease: Causes and Treatment," *Molecules*, vol. 25,

no. 24, p. 5789, Dec. 2020.

[2] I. Arevalo-Rodriguez *et al.*, "Mini-Mental State Examination (MMSE) for the detection of Alzheimer's disease and other dementias in people with mild cognitive impairment (MCI)," *The Cochrane Database of Systematic Reviews*, vol. 2015, no. 3, p. CD010783, Mar. 2015.

[3] C. Carnero-Pardo, "Should the mini-mental state examination be retired?" *Neurologia (Barcelona, Spain)*, vol. 29, no. 8, pp. 473–481, Oct. 2014.

[4] S. de la Fuente Garcia *et al.*, "Artificial Intelligence, Speech, and Language Processing Approaches to Monitoring Alzheimer's Disease: A Systematic Review," *Journal of Alzheimer's Disease*, vol. 78, no. 4, pp. 1547–1574, Dec. 2020.

[5] J. Becker *et al.*, "The natural history of Alzheimer's disease. Description of study cohort and accuracy of diagnosis," *Archives of Neurology*, vol. 51, no. 6, pp. 585–594, Jun. 1994.

[6] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's Dementia Recognition Through Spontaneous Speech: The ADReSS Challenge," in *Proc. Interspeech 2020*, 2020, pp. 2172–2176.

[7] ——, "Detecting Cognitive Decline Using Speech Only: The ADReSSo Challenge," in *Proc. Interspeech 2021*, 2021, pp. 3780–3784.

[8] R. Pappagari *et al.*, "Using state of the art speaker recognition and natural language processing technologies to detect alzheimer's disease and assess its severity." in *INTERSPEECH*, 2020, pp. 2177–2181.

[9] A. Meghanani, C. Anoop, and A. Ramakrishnan, "An exploration of log-mel spectrogram and mfcc features for alzheimer's dementia recognition from spontaneous speech," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 670–677.

[10] U. Sarawgi *et al.*, "Multimodal Inductive Transfer Learning for Detection of Alzheimer's Dementia and its Severity," in *Proc. Interspeech 2020*, 2020, pp. 2212–2216.

[11] N. Cummins *et al.*, "A Comparison of Acoustic and Linguistics Methodologies for Alzheimer's Dementia Recognition," in *Proc. Interspeech 2020*, 2020, pp. 2182–2186.

[12] R. Haulcy and J. Glass, "Classifying alzheimer's disease using audio and text-based representations of speech," *Frontiers in Psychology*, vol. 11, p. 624137, 2021.

[13] Y. Pan *et al.*, "Using the outputs of different automatic speech recognition paradigms for acoustic-and bert-based alzheimer's dementia detection through spontaneous speech." in *Interspeech*, 2021, pp. 3810–3814.

[14] A. Balagopalan *et al.*, "To BERT or not to BERT: Comparing Speech and Language-Based Approaches for Alzheimer's Disease Detection," in *Interspeech 2020*. ISCA, Oct. 2020, pp. 2167–2171.

[15] M. S. S. Syed *et al.*, "Automated Screening for Alzheimer's Dementia Through Spontaneous Speech," in *Interspeech 2020*. ISCA, Oct. 2020, pp. 2222–2226.

[16] J. Yuan *et al.*, "Disfluencies and Fine-Tuning Pre-Trained Language Models for Detection of Alzheimer's Disease," in *Interspeech 2020*. ISCA, Oct. 2020, pp. 2162–2166.

[17] Y.-W. Chien, S.-Y. Hong, W.-T. Cheah, L.-H. Yao, Y.-L. Chang, and L.-C. Fu, "An automatic assessment system for alzheimer's disease based on speech using feature sequence generator and recurrent neural network," *Scientific Reports*, vol. 9, no. 1, pp. 1–10, 2019.

[18] C. Sanz *et al.*, "Automated text-level semantic markers of alzheimer's disease," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 14, no. 1, p. e12276, 2022.

[19] G. Gosztolya *et al.*, "Cross-lingual detection of mild cognitive impairment based on temporal parameters of spontaneous speech," *Computer Speech & Language*, vol. 69, p. 101215, 2021.

[20] P. A. Pérez-Toro *et al.*, "Alzheimer's Detection from English to Spanish Using Acoustic and Linguistic Embeddings," in *Proc. Interspeech 2022*, 2022, pp. 2483–2487.

[21] S. Luz *et al.*, "Multilingual alzheimer's dementia recognition through spontaneous speech: a signal processing grand challenge," 2023. [Online]. Available: https://arxiv.org/abs/2301.05562

[22] O. Ivanova *et al.*, "Discriminating speech traits of alzheimer's disease assessed through a corpus of reading task for spanish language," *Computer Speech & Language*, vol. 73, p. 101341, 2022.

[23] H. Bredin *et al.*, "pyannote.audio: neural building blocks for speaker diarization," in *ICASSP 2020*, 2020.

[24] T. Giannakopoulos, "pyaudioanalysis: An open-source python library for audio signal analysis," *PloS one*, vol. 10, no. 12, p. e0144610, 2015.

[25] J. Shor and S. Venugopalan, "TRILLsson: Distilled Universal Paralinguistic Speech Representations," in *Proc. Interspeech 2022*, 2022, pp. 356–360.

[26] Y. Gong *et al.*, "AST: Audio Spectrogram Transformer," in *Proc. Interspeech 2021*, 2021, pp. 571–575.

[27] A. Radford *et al.*, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022.

[28] M. Honnibal *et al.*, "spaCy: Industrial-strength Natural Language Processing in Python," 2020.

[29] F. Eyben *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.