INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Using state of the art speaker recognition and natural language processing technologies to detect Alzheimer's disease and assess its severity

Raghavendra Pappagari, Jaejin Cho, Laureano Moro-Velazquez, Najim Dehak

Center for Language Speech Processing, Johns Hopkins University, Baltimore, MD, USA
{rpappag1,jcho52,laureano,ndehak3}@jhu.edu

Abstract

In this study, we analyze the use of state-of-the-art technologies for speaker recognition and natural language processing to detect Alzheimer's Disease (AD) and to assess its severity predicting Mini-mental status evaluation (MMSE) scores. With these purposes, we study the use of speech signals and transcriptions. Our work focuses on the adaptation of state-of-the-art models for both modalities individually and together to examine its complementarity. We used x-vectors to characterize speech signals and pre-trained BERT models to process human transcriptions with different back-ends in AD diagnosis and assessment. We evaluated features based on silence segments of the audio files as a complement to x-vectors. We trained and evaluated our systems in the Interspeech 2020 ADReSS challenge dataset, containing 78 AD patients and 78 sex and age-matched controls. Our results indicate that the fusion of scores obtained from the acoustic and the transcript-based models provides the best detection and assessment results, suggesting that individual models for two modalities contain complementary information. The addition of the silence-related features improved the fusion system even further. A separate analysis of the models suggests that transcript-based models provide better results than acoustic models in the detection task but similar results in the MMSE prediction task.

1. Introduction

Alzheimers Disease (AD) is the most common cause of dementia and the most prevalent neurodegenerative condition. Its impact on the multiple aspects of society is rising due to the aging of the worldwide population [1]. While two of the most typical signs of AD are memory and cognitive decline, the literature suggests that language impairment is also a common sign that can be employed to support diagnosis and assessment of the severity of the disease, given that speech and language production can provide information about the cognitive status of a person and other aspects related to brain damage. Although the human evaluation of speech and language can be used to diagnose and assess patients in the clinical setting, that type of evaluation does not allow an objective quantitative analysis and reliable repeatability. To this respect, the use of speech recognition and Natural Language Processing (NLP) techniques can deliver new precision medicine tools that will provide objective measures and biomarkers. This will allow faster diagnosis and assessment in a non-invasive and cost-effective manner.

Although the influence of AD in speech and language is diverse and subject-dependent, the literature suggests some common signs such as progressive, logopenic or anomic aphasia [2, 3, 4] (communication and word retrieval impairment, phone substitution) and apraxia of speech [5] (articulatory impairment.) Therefore, several studies indicate that both phonetic-motor signs (related to apraxia) and phonological-linguistic manifestations (related to aphasia and anomia) can be found in cohorts of AD patients [5]. Depending on the patient, the apraxic or aphasic manifestations can be prevalent, suggesting that both acoustic and linguistic analyses are advisable in systems employing speech technologies automatically to detect AD or assess its severity.

In this respect, the combination of acoustic and linguistic features within machine learning based-approaches to automatically detect AD in recordings obtained from the DementiaBank corpus has already been analyzed [6], obtaining 81% cross-validation accuracy. Other studies providing similar results suggest that linguistic features provide higher accuracy than acoustic features in detecting AD [7]. However, the combination of both types of features yields better results than when using these features separately, suggesting that these features are complementary [7]. Additionally, accuracies over 80% have been reported when employing word and silence rates obtained with Voice Activity Detection (VAD) systems and transcripts [8]. Moreover, some linguistic features indicative of lexical diversity such as word frequency, percentage of content words, pronoun ratio or type-token ratio among others have shown a high correlation with Mini-Mental Status Examination (MMSE) in AD patients [9], suggesting that patient's morphosyntactic impairments can be automatically analyzed and employed for severity assessment.

Although the literature includes a fair amount of studies employing acoustic and linguistic features [6, 7, 8, 9, 10, 11, 12] for the automatic detection and assessment of AD, to our knowledge no study analyzes the use of speaker recognition and NLP technologies such as x-vectors [13] and Bidirectional Encoder Representations from Transformers (BERT) [14]. These techniques have become the state-of-the-art in speech technologies, and its acoustic and linguistic characterization properties have been exploited in multiple scenarios such as Parkinson's Disease (PD) detection [15], emotion recognition [16], sentiment analysis [17] or question answering [14], among others.

Consequently, this study aims to analyze the use of these two Deep Neural Networks (DNN)-based techniques, x-vectors and BERT, in AD detection and MMSE prediction scenarios.

2. ADReSS Challenge Dataset

The ADReSS Challenge dataset [18] contains two subsets with speech and transcriptions from speakers with and without AD: the *training* and the *evaluation subsets*. In this study, the *training subset* was used to perform cross-validation and to train models to be evaluated with the *evaluation subset*.

The *training subset* includes two groups of speakers: those diagnosed with AD (AD group) and the age- and sex-matched control speakers (CC group). Each group is composed of 24 male and 30 female participants. Data in both groups contain one audio recording per participant, recorded at 44100 Hz and

http://dx.doi.org/10.21437/Interspeech.2020-2587

with an average length of 72.10 s, demographic information, full transcript, and MMSE score. In our experiments, we down-sampled the recordings according to the models we used, as explained in later sections.

The *evaluation subset* comprises 11 male and 13 female participants in each group, while the age distribution is the same over the two groups. The average session length is 82.51 s. Challenge participants do not have information about AD diagnosis or MMSE assessment for these speakers.

3. Experimental Setup

In this study, we employed two main models to detect AD and predict MMSE from speech. The first model or acoustic model is based on the use of acoustic aspects of speech and employs a speaker characterization technique, i.e., x-vectors and two different back-ends: Probabilistic Linear Discriminant Analysis (PLDA) for detection and Support Vector Regression (SVR) for MMSE prediction. The x-vectors were complemented with heuristic features obtained from the analysis of the silence and pause segments from the speech signal. The second model or transcript-based model is a BERT model that utilizes linguistic contents to detect AD subjects and predict MMSE. We hypothesize that the transcript-based model provides complementary information to the acoustic model. Finally, scores from the two approaches were fused using a Gradient Boosting Regressor (GBR) or averaging, depending on the task.

Moreover, we differentiate two types of results:

- Cross-validation results: obtained training and testing with the *training subset*, using a 10-fold scheme where class and age distributions were consistent over the folds. The cross-validation was done speaker-independently since the dataset has only one session recorded per participant.
- Evaluation results: obtained by testing the models trained with the *training subset* on the *evaluation subset*.

3.1. Acoustic model

3.1.1. x-vectors

To model the speakers' articulatory, prosodic and phonatory characteristics included in the dataset, we employ representation obtained with an x-vector model trained for speaker recognition. An x-vector model is a deep neural network that generates one single vector or embedding per utterance, characterizing the speaker. Although the technique is considered the current state-of-the-art for speaker recognition, several studies suggest that these embeddings also contain information related to emotion, speaking rate, gender [16, 19] and other articulatory, phonatory and prosodic information that can be used to characterize neurological diseases, as Parkinson's Disease [15]. In general terms, an x-vector model contains three main parts: an encoder network to extract frame-level representation from MFCC, a global temporal pooling layer to produce the embedding (x-vector), and a feed-forward classification network to produce speaker class posteriors. Once the model training is done, only the first two parts are used while the last part is discarded. In our case, the three parts consisted of a factorized time delay network encoder (F-TDNN), mean plus standard deviation pooling, and two feed-forward layers, respectively, as detailed in a previous study [15]. Differentiation process between AD and CC speakers followed the same setup as the one explained in the cited study:

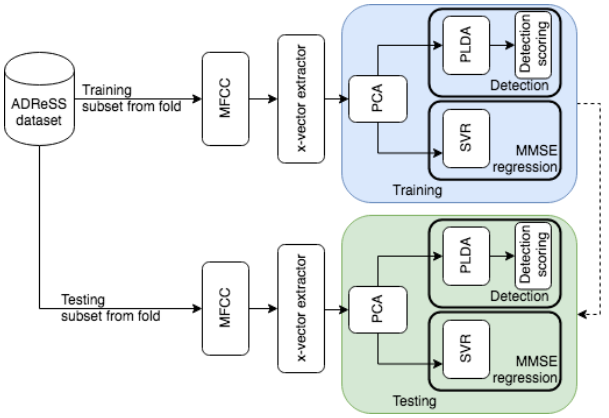


Figure 1: *Diagram of the acoustic model methodology. In cross-validation stage, models obtained with the training folds are used for testing with their respective testing folds. In evaluation stage, the whole training dataset is employed for training while the evaluation dataset is used for testing*

- First, all speech signals were normalized, low-pass filtered and re-sampled to 16 kHz.
- Then, we extracted MFCC features (40 coefficients, frame length of 25 ms with frame shift 10 ms)
- Silence segments were removed employing the standard VAD from Kaldi [20].
- MFCC features were used to extract one x-vector (dimension 512) for each speech recording using an x-vector model trained with VoxCeleb 1 and 2 corpora [21, 22] in Kaldi with sampling frequency 16 kHz.
- At each cross-validation iteration, all the x-vectors from the training folds were employed to train a Principal Component Analysis (PCA) model that was applied to the x-vectors from the training and testing folds in the cross-validation stage.
- For AD detection, x-vector PCA-transformed coefficients from the training folds were used to train a PLDA classifier to differentiate between AD and CC speakers. In the classifier, a likelihood ratio per speech recording is calculated considering two classes (AD and CC) which is employed in scoring to take the decision. The scoring threshold is set to the equal error rate point obtained with the log-likelihoods from the training folds x-vector-PCA coefficients.
- Similarly, for MMSE prediction, we trained and evaluated a linear SVR on the x-vector PCA-transformed coefficients.

Fig. 1 includes a diagram of the described process. To get the best PCA and PLDA models for evaluation on the *evaluation subset*, the whole ADReSS *training subset* was used.

3.1.2. Silence features

To complement the x-vectors characterization, which is data-driven, we also extracted 4-dimensional heuristic features based on the Kaldi energy-based VAD algorithm. Our goal was to characterize the presence of silences in the recordings. The four features are:

- Silence rate (the number of silence regions divided by the recording length)
- Ratio of silence to speech duration

- Mean and standard deviation of the duration of silence regions

We only considered silence regions that were longer than 150 ms. Also, we removed the silences at the start and end of the recordings when these existed. We considered these features since previous studies suggest that silence-related features can help to characterize aphasia and apraxia associated with AD [8]. We used these features in two different manners in this study:

- As single features for PLDA and SVR model training to examine the discrimination capabilities of these features.
- Appended to the x-vector PCA-transformed coefficients, which we denominate *Acoustic model with silence features* scheme. This allows us to observe the complementarity between x-vectors and silence features.

3.2. Transcript-based model

To model the linguistic-phonological manifestations of AD on speech, we employed a BERT model [14] on the spoken transcripts, which has shown state-of-the-art performances in several NLP applications such as question answering, natural language inference, named entity recognition, sentence, and word prediction, among many others. We chose BERT for two reasons: 1) the embeddings obtained from this model act as general text representation and, 2) previous studies reported good results from fine-tuned BERT models for multiple tasks. Two examples are depression detection [23] or sentiment analysis [17].

BERT is a pre-trained language model trained to predict masked words of a sentence and the next sentence. The BERT architecture mainly consists of self-attention layers and feed-forward layers. In general, a pre-trained BERT model is adapted to a down-stream task by fine-tuning the pre-trained parameters with the minimal number of newly introduced parameters for the task [14]. We adapted BERT to our tasks (AD detection and MMSE prediction) in a similar way:

- We replaced the last layer of the BERT model with a task-specific layer: a linear layer having two neurons with a softmax activation function for AD detection or a linear layer having 1 neuron with linear activation function for MMSE prediction.
- We fine-tuned the entire pre-trained model using our data to minimize cross-entropy loss for AD detection or mean square error for MMSE prediction.

The inputs of the model were tokens from the transcript that were tokenized into sub-words using WordPiece tokenizer [24]. These inputs were processed through multiple self-attention and feed-forward layers to obtain embeddings for each sub-word in the penultimate layer. Then, the sequence of sub-word embeddings was pooled to pass through a linear layer to obtain the prediction.

For each iteration of the cross-validation experiments, 9 folds from the *training subset* were employed for BERT fine-tuning and the remaining fold for testing. We used early stopping criterion to stop training the model and trained for 5 epochs.

3.3. Fusion

In this section, we describe our methodology for fusing acoustic and transcript model scores. For the AD detection task, we first obtained the scores from acoustic and transcript-based models for all utterances from the testing folds during the cross-validation stage. Then, we employed these predictions in a cross-validation scheme to train and test the fusion of the scores using a GBR model [25], which provided the cross-validation results. To obtain the *evaluation subset* predictions, we employed the scores from the whole *training subset* to train a final fusion GBR model that was used to perform the fusion of scores coming from the acoustic and transcript-based models for the challenge evaluation. For MMSE prediction, we followed a similar procedure but simply averaged the scores from the different models instead of using a GBR.

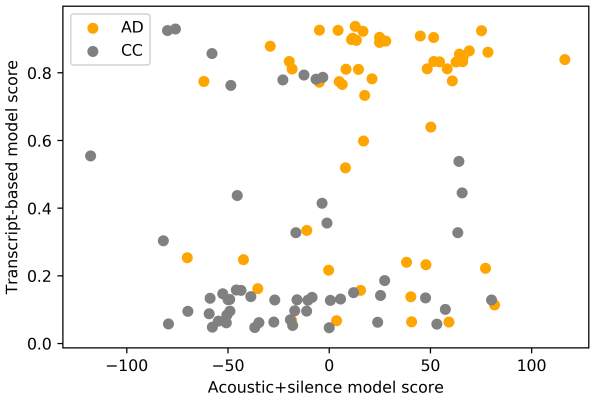


Figure 2: *Score scatterplot for AD and CC speakers in detection task considering the transcript-based model scores (that range between 0 and 1) and the log-likelihood ratio obtained with the PLDA classifier for the acoustic+silence model. Each dot represent one subject.*

4. Results and Discussion

In this section, we present our results on both AD detection and MMSE prediction tasks. For evaluation metrics, we used the same metrics as proposed in [18], namely, accuracy, precision, recall, and F1 score for detection and Root Mean Square Error (RMSE) for MMSE prediction. For simplicity, in cross-validation results (10 folds) we only report accuracy and RMSE.

4.1. Cross-validation results

Table 1 presents the cross-validation results with the proposed models for AD detection and MMSE prediction tasks. From the comparison of acoustic and transcript models, we can observe that the transcript-based model performed better than the acoustic model for AD detection but worse in MMSE prediction. The use of silence features alone did not provide high accuracy to differentiate between AD and CC groups. However, when we concatenated silence features with x-vectors PCA-transformed coefficients, denoted as Acoustic+silence in Table 1, we obtained an absolute 2.4% improvement in AD detection accuracy compared to using acoustic features alone, implying that acoustic and silence features may have complementary information. For MMSE prediction task, we obtained a small improvement in RMSE value after concatenation (0.03, absolute).

We further fused acoustic and transcript-based model scores to exploit their complementary information. The fusion model showed 79.2% accuracy and 5.93 RMSE, which indicates a 0.5% and 0.31 improvement compared to the best individual model, respectively. Thus, results suggest that score fu-

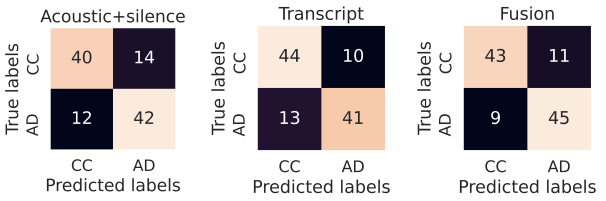
sion provides improvements in both AD detection and MMSE prediction. In the same sense, the fusion of Acoustic+silence and transcript models scores yielded 81.44% AD accuracy and 5.91 RMSE for MMSE prediction, the best cross-validation results.

A scatterplot of detection scores per subject is shown in Figure 2. The figure indicates that in the detection task, the transcript-based analysis is more informative for some speakers, while the acoustic signal analysis is so for some of the others. We can observe that for the majority of subjects, the scores from the two types of models help to cluster the two groups of speakers in the bottom left (CC) and upper right (AD) parts of the score bi-dimensional space, suggesting that both acoustic signal and transcripts contain cues to detect AD. Nevertheless, a few subjects have opposite results in different models, showing a high score from the transcript-based model but a low score from the acoustic model and vice versa. This indicates that different models can provide complementary information.

Figure 3 shows the confusion matrices of the models with the best cross-validation results. We can observe that the models are not biased to any single class, i.e., the recall for each class, AD and CC, is similar. Improvement in the fusion model is reflected with higher diagonal values and lower off-diagonal values in general, compared to the two individual models.

Table 1: *Cross-validation (CV) results for AD detection and MMSE precition tasks. Best results are marked in bold.*

Models	Detection CV accuracy (%)	Prediction RMSE
Acoustic	73.21	6.24
Silence	51.20	8.05
Acoustic+silence	75.93	6.21
Transcript	78.70	6.52
Acoustic & Transcript	79.20	5.93
Acoustic+silence & Transcript	**81.48**	**5.91**



(a) *(b)* *(c)*

Figure 3: *Confusion matrices for the detection tasks using (a) Acoustic model with silence features, (b) Transcript model, (c) Fusion of Acoustic model with silence features and transcript model.*

4.1.1. Evaluation results

Results for the *evaluation subset* were obtained from the submission of our model predictions to the ADReSS challenge organizers. Table 2 shows the evaluation results of our models in AD detection and MMSE prediction tasks. Baseline results are based on the use of the ComParE 2013 feature set [26] and a linear discriminant analysis classifier (for detection) and MRGC features [27] with decision trees (for MMSE prediction.) These baseline results were provided by the ADReSS challenge organizers [18]. We observed that four of our four models outperformed the baseline in the detection task by significant margins, and all of them provided a better RMSE than the baseline. The model comparison showed similar trends in accuracy on the evaluation and cross-validation results, but the overall accuracy was lower in the evaluation than the cross-validation. For MMSE prediction, all RMSE values are lower in the evaluation experiments than in the cross-validation. The model providing the best accuracy was the score-level fusion of acoustic and transcript models with 75% accuracy. When silence features were also used, we obtained the best MMSE prediction results, 5.32 RMSE.

We observed that the acoustic model performance in the *evaluation subset* is much lower than its correspondent cross-validation accuracy, suggesting that the acoustic models might have been overfitted to the *training subset*. We observed the same trends and conclusions from model comparison in cross-validation and evaluation experiments in Tables 1 and 2, as the complementarity between transcript and acoustic models.

Table 2: *ADReSS challenge evaluation results for the detection and prediction tasks. Best results are marked in bold.*

Models	Class	Detection Prec./Rec.	F1	Accuracy (%)	Prediction RMSE
Baseline	CC	0.67/0.50	0.57	62.50	6.14
	AD	0.60/0.75	0.67		
Acoustic	CC	0.61/0.45	0.52	58.00	6.08
	AD	0.57/0.71	0.63		
Acoustic + silence	CC	0.64/**0.75**	0.69	66.70	5.97
	AD	0.70/0.58	0.63		
Transcript	CC	0.79/0.63	0.7	72.92	5.86
	AD	0.69/0.83	0.75		
Acoustic & Transcript	CC	**0.83**/0.63	**0.71**	**75.00**	5.37
	AD	**0.70/0.88**	**0.78**		
Acoustic + silence & Transcript	CC	0.79/0.62	0.70	72.92	**5.32**
	AD	0.69/0.83	0.75		

5. Conclusions and future work

This study presents different approaches to automatically detect AD and predict MMSE from the speech signal and its associated transcript, based on the acoustic characterization of the speech signal and the transcript-based modeling employing DNN. The employed x-vectors and BERT are considered the current state-of-the-art techniques in speaker recognition and NLP, respectively. Our results suggest that transcription-based models provide better results in detection and prediction tasks than acoustic models. At the same time, information about the silences present in the recording improves accuracy for acoustic modeling. The best results in cross-validation and evaluation stages are obtained with the fusion of the scores provided by both the acoustic and transcript-based models.

In future work, we will explore the x-vector adaptation by fine-tuning the extractor [16] for the AD/CC detection task. Also, we will explore the use of automatic speech recognition systems to obtain the speech transcription and compare results with human transcription. Lastly, we will explore the use of BioBERT [28] and other transformer-based architectures for the detection and assessment of AD.

6. References

[1] K. B. Rajan, J. Weuve, L. L. Barnes, R. S. Wilson, and D. A. Evans, "Prevalence and incidence of clinically diagnosed alzheimer's disease dementia from 1994 to 2012 in a population study," *Alzheimer's & Dementia*, vol. 15, no. 1, pp. 1–7, 2019.

[2] J. D. Rohrer, M. N. Rossor, and J. D. Warren, "Alzheimer's pathology in primary progressive aphasia," *Neurobiology of aging*, vol. 33, no. 4, pp. 744–752, 2012.

[3] S. Ahmed, C. A. de Jager, A.-M. F. Haigh, and P. Garrard, "Logopenic aphasia in alzheimer's disease: clinical variant or clinical feature?" *J Neurol Neurosurg Psychiatry*, vol. 83, no. 11, pp. 1056–1062, 2012.

[4] S. M. Harnish, "Anomia and anomic aphasia: Implications for lexical processing." *The Oxford Handbook of Aphasia and Language Disorders*, 2018.

[5] E. Rochon, C. Leonard, and M. Goral, "Speech and language production in alzheimer's disease," *Aphasiology*, vol. 32, no. 1, pp. 1–3, 2018.

[6] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify alzheimer's disease in narrative speech," *Journal of Alzheimer's Disease*, vol. 49, no. 2, pp. 407–422, 2016.

[7] G. Gosztolya, V. Vincze, L. Tóth, M. Pákáski, J. Kálmán, and I. Hoffmann, "Identifying mild cognitive impairment and mild alzheimer's disease based on spontaneous speech using asr and linguistic features," *Computer Speech & Language*, vol. 53, pp. 181–197, 2019.

[8] J. Weiner, C. Herff, and T. Schultz, "Speech-based detection of alzheimer's disease in conversational german." in *INTERSPEECH*, 2016, pp. 1938–1942.

[9] G. Kavé and A. Dassa, "Severity of alzheimer's disease and language features in picture descriptions," *Aphasiology*, vol. 32, no. 1, pp. 27–40, 2018.

[10] S. Luz, "Longitudinal monitoring and detection of alzheimer's type dementia from spontaneous speech data," in *2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*, 2017, pp. 45–46.

[11] S. Luz, S. de la Fuente, and P. Albert, "A method for analysis of patient speech in dialogue for dementia detection," *Proc. of the LREC 2018 Workshop "Resources and ProcessIng of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric impairments (RaPID-2)",*, 2018.

[12] S. Karlekar, T. Niu, and M. Bansal, "Detecting linguistic characteristics of alzheimer's dementia by interpreting neural models," *arXiv preprint arXiv:1804.06440*, 2018.

[13] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *ICASSP*, 2018, pp. 5329–5333.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[15] L. Moro-Velazquez, J. Villalba, and N. Dehak, "Using x-vectors to automatically detect parkinson's disease from speech," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 1155–1159.

[16] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, "x-vectors meet emotions: A study on dependencies between emotion and speaker recognition," *arXiv preprint arXiv:2002.05039*, 2020.

[17] S. Pei, L. Wang, T. Shen, and Z. Ning, "Da-bert: Enhancing part-of-speech tagging of aspect sentiment analysis using bert," in *International Symposium on Advanced Parallel Processing Technologies*. Springer, 2019, pp. 86–95.

[18] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The ADReSS Challenge," in *Proceedings of INTERSPEECH 2020*, Shanghai, China, 2020. [Online]. Available: https://arxiv.org/abs/2004.06833

[19] D. Raj, D. Snyder, D. Povey, and S. Khudanpur, "Probing the information encoded in x-vectors," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2019.

[20] D. Povey, A. Ghoshal, and G. Boulianne, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, 2011.

[21] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," in *Interspeech*, aug 2017, pp. 2616–2620.

[22] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Interspeech*, 2018, pp. 1086–1090.

[23] M. Rodrigues Makiuchi, T. Warnita, K. Uto, and K. Shinoda, "Multimodal fusion of bert-cnn and gated cnn representations for depression detection," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 2019, pp. 55–63.

[24] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado *et al.*, "Google's multilingual neural machine translation system: Enabling zero-shot translation," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 339–351, 2017.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[26] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 835–838.

[27] J. Chen, Y. Wang, and D. Wang, "A feature study for classification-based speech separation at low signal-to-noise ratios," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1993–2002, 2014.

[28] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.