A COMPARATIVE STUDY OF ACOUSTIC AND LINGUISTIC FEATURES CLASSIFICATION FOR ALZHEIMER'S DISEASE DETECTION

Jinchao Li[1], *Jianwei Yu*[1], *Zi Ye*[1], *Simon Wong*[1], *Manwai Mak*[2], *Brian Mak*[3], *Xunying Liu*[1], *Helen Meng*[1]

[1]The Chinese University of Hong Kong, [2]The Hong Kong Polytechnic University
[3]The Hong Kong University of Science and Technology

[1]{jcli,jwyu,zye,khwong,xyliu,hmmeng}@se.cuhk.edu.hk, [2]enmwmak@polyu.edu.hk, [3]mak@cse.ust.hk

ABSTRACT

With the global population ageing rapidly, Alzheimer's disease (AD) is particularly prominent in older adults, which has an insidious onset followed by gradual, irreversible deterioration in cognitive domains (memory, communication, etc). Thus the detection of Alzheimer's disease is crucial for timely intervention to slow down disease progression. This paper presents a comparative study of different acoustic and linguistic features for the AD detection using various classifiers. Experimental results on ADReSS dataset reflect that the proposed models using ComParE, X-vector, Linguistics, TF-IDF and BERT features are able to detect AD with high accuracy and sensitivity, and are comparable with the state-of-the-art results reported. While most previous work used manual transcripts, our results also indicate that similar or even better performance could be obtained using automatically recognized transcripts over manually collected ones. This work achieves accuracy scores at 0.67 for acoustic features and 0.88 for linguistic features on either manual or ASR transcripts on the ADReSS Challenge[1] test set.

Index Terms— Alzheimer's Disease detection, ADReSS, features, ASR

1. INTRODUCTION

Alzheimer's disease (AD), a major kind of neurocognitive disease (also called dementia), is characterized by clear decline of cognitive functioning, including memory, language, thinking and behavior [1]. It has been estimated in 2019 that AD affects over 50 million people globally, and worse still, the disease has insidious onset with irreversible deterioration and unknown therapy nowadays [2]. Therefore, the detection of AD is crucial for the timely intervention and to decelerate progression. Conventional methods for AD detection are mainly based on clinical tests for cognitive decline and independence in everyday activities [3]. However, these diagnostic processes are constrained due to time requirements and accessibility to resources. Since spoken language is an easily captured signal that can reflect the speaker's cognitive abilities, researchers have been motivated to investigate the use of speech (the acoustic signal) and language (words and sentences) features as biomarkers for AD detection [4, 5, 6, 7].

Feature selection is an important first step to search for speech and language features that can help distinguish between participants in the dataset who are healthy and participants who have AD. Weiner et al [6] proposed a screening method and ranked the importance of hand-selected acoustic and linguistic features for longitudinal AD prediction. Most of the top-ranking features are linguistic, such as parts-of-speech and word categories, while the acoustic features, such as I-vector and pause-based features, did not fare as well in AD detection In addition to the use of hand-crafted features, deep neural networks and transfer learning models have shown promising performance in recent years as feature extractors for AD detection in recent years [8, 9, 10]. In utilizing the speech signal, much of the existing research efforts are manual transcriptions, which is costly to obtain. This paper aims to further explore the use of acoustic features, especially in the use of automatic transcriptions in deriving linguistic features for AD detection .

This paper presents a comparative study of comprehensive acoustic and linguistic features for Alzheimer's Disease detection using different classifiers. The features covered include ComParE, X-vector derived from the acoustics signal, and linguistic features, TF-IDF, BERT derived from the manually/automatically transcribed text. This involves a comparison in detection performance between manual and automatic transcriptions. The classifiers covered include Linear Discriminant Analysis (LDA), Support Vector Machine (SVM) and Attention-based Long Short-Term Memory Recurrent Neural Network (AT-LSTM) [11]. We adopt Pearson's correlation test for feature selection and principal components analysis (PCA) for dimensionality reduction [12]. Evaluation of these features and models are based on the balanced benchmark ADReSS dataset [13], which is a subset of Pitt Corpus in DementiaBank database [14].

The main contributions of this paper are summarized below. We compare and improve the performance of speech and language features on different classifiers, followed by further analysis to uncover the features that are highly related to Alzheimer's Disease (AD). Our experimental results also indicate that the AD detection performance based on textual features derived from ASR (i.e. Automatic Speech Recognition) transcripts can be as good as manual transcripts.

In the next section, we propose the overall detection system, and describe the dataset, AD-related features and classifiers used in our work. Following that, we introduce the experimental details and results of the proposed detection systems. We compare the performance of the proposed systems with the published benchmarks in [13], and highlight the results on the proposed ASR-based detection system. Finally, in Section 4, we discuss the performance of the features and classifiers, identify the features that are highly related to AD, and present our conclusions and future directions of research.

2. METHODOLOGY

In this section, we first introduce our proposed AD detection system, and then describe the ADReSS dataset, the different types of features, and the various classifiers LDA, SVM and AT-LSTM used in our work.

[1]The ADReSS Challenge: http://www.homepages.ed.ac.uk/sluzfil/ADReSS/

2.1. Overall AD Detection System

The overall process for AD detection mainly consists of data preparation, feature engineering and classification, as shown in Fig. 1. The input to the detection system are elderly speech recordings from the ADReSS dataset, together with their corresponding manual transcription. To achieve automatic detection, we also use automatic transcriptions of the speech recordings generated by an ASR system. We extract different types of features from the audio or text, and select AD-related features with the Pearson's correlation test. To avoid over-fitting, we also perform PCA to reduce the feature space. Then we feed the processed features into the classifiers for training, and finally the trained systems can perform AD detection.

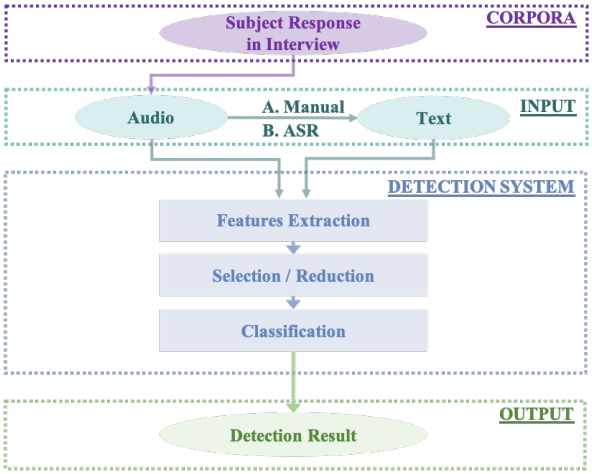

Fig. 1. Overall process for AD detection.

2.2. Dataset

The dataset we used in this work is the ADReSS Challenge 2020 dataset, which is a selected part of Pitt Corpus in the DementiaBank database. The dataset consists of 156 speech samples and associated transcripts from non-AD (35 male, 43 female) and AD (35 male, 43 female) English-speaking participants for the Cookie Theft picture description task, and is divided into standard train (108 participants, about 2 hours) and test (48 participants, about 1 hour) sets that balanced for age, gender and disease condition.

2.3. Feature Engineering

2.3.1. Acoustic & Linguistic Features

Speech and language impairments caused by Alzheimer's disease affect temporal changes in speech, verbal fluency, word finding and word retrieval abilities in language [4, 6, 15]. Based on these findings, we explore the ComParE [16] and Linguistics [17] feature sets, which include various descriptors for speech (the acoustic signal) and language (words and sentences) respectively. ComParE is a feature set that consists of generic acoustic emotion descriptors and their statistical functionals, including temporal features, voicing related low-level descriptors (LLDs), etc. The Linguistics features set is a set of language outcome measures, including lengths of utterances, type-token ratios, statistics of parts-of-speech (POS), etc.

Based on the differences in lexical richness and fluency between AD and non-AD participants, we use the Term Frequency-Inverse Document Frequency (TF-IDF) vector as the representation of words [18]. Moreover, we also explore the use of representations based on transfer learning, such as X-vector [19] and BERT [20]. The X-vector captures long-term speaker characteristics for speech, and BERT is a kind of deep bidirectional representation for text. Both of them have achieved state-of-the-art results on a wide variety of speech and language tasks. We adopt the pre-trained models for X-vector and BERT as mentioned respectively in [19] and [20], and extract the corresponding features from the ADReSS dataset.

2.3.2. ASR-derived Linguistic Features

We use the ASR transcripts [21] to derive linguistic features including TF-IDF vector and BERT embeddings. The ASR systems are trained on original Pitt Corpus data that excluded ADReSS test participants, and testing on the ADReSS test set gave word error rates (WER) 44.89% (Sys. 4) and 33.17% (Sys. 10) for the respective participant parts. The methods of extracting the TF-IDF vectors and BERT embeddings from the manual and ASR transcripts are the same.

2.3.3. Features Selection / Reduction

We perform Pearson's correlation test to select AD-related features, i.e., select the features that have high correlation coefficient with the occurrence of the disease. To avoid over-fitting, we perform PCA to reduce the dimensionality of the feature space, i.e., we first fit the parameters of PCA with the features from training set, and transform the features from all data with the fitted PCA model.

2.4. Models

We formulate AD detection as a binary classification problem, i.e. classifying AD and non-AD participants. We refer to the LDA classifier as baseline, as reported in [13], and then proposed SVM and AT-LSTM based on supervised learning.

2.4.1. LDA

Suppose the two classes of observations have means $\mu_0, \mu_1 \in \mathcal{R}^M$ and covariances $\Sigma_0, \Sigma_1 \in \mathcal{R}^{M\times M}$. The objective of LDA is to find a projection direction $w \in \mathcal{R}^M$, such that the ratio between the variance of projected points between to within classes is maximized, i.e.:

$$\arg\max_w \frac{w^T(\mu_0-\mu_1)(\mu_0-\mu_1)^T w}{w^T(\Sigma_0+\Sigma_1)w}. \tag{1}$$

Assuming that the singular value decomposition (SVD) of matrix $(\Sigma_0+\Sigma_1)$ is $U\Sigma V^T$, then Eq. 1 can be solved as:

$$w = V\Sigma^{-1}U^T(\mu_0-\mu_1). \tag{2}$$

2.4.2. SVM

Suppose the training set of N points are $\{(x_i, y_i), x_i \in \mathcal{R}^M\}_{i=1}^N$. The SVM classifier tries to find the "maximum-margin hyperplane" that divides the two different groups of points, represented by $w^T x + b = 0$. Then the objective of SVM is to minimize the aggregate distance between the maximum-margin hyperplane and the support vectors of AD and non-AD classes, i.e.:

$$\arg\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\mathcal{L}(y_i(w^T\phi(x_i)+b)), \tag{3}$$

where C is the regularization parameter for soft margin, $\mathcal{L}(*)$ is the surrogate loss and $\phi(*)$ is the kernel function. In this work, we set $C = 1$, surrogate loss $\mathcal{L}(z) = \max(0, 1 - z)$ and kernel function as linear kernel. Then Eq. 3 can be solved by dual programming method.

2.4.3. AT-LSTM

In addition to the LDA and SVM classifiers, we also propose to use AT-LSTM for AD versus non-AD classification. The common LSTM has feedback connections and is composed of the memory cell, input, output and forget gates, which can capture time domain representations. The attention mechanism can capture the key parts of feature vector in response to a given aspect. Fig. 2 represents the AT-LSTM architecture used in our work.

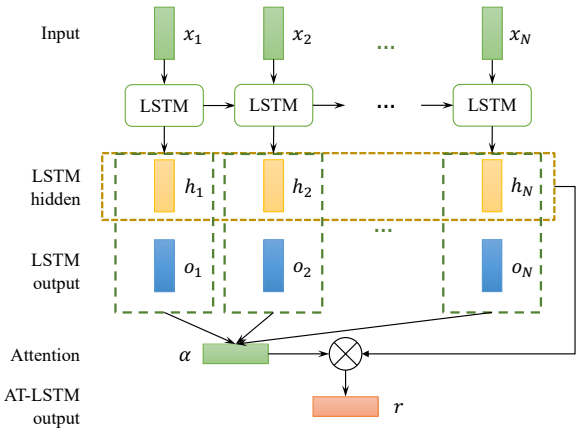

Fig. 2. The architecture of Attention-based LSTM module.

Suppose the input features $X = [x_1, x_2, \ldots, x_N] \in \mathcal{R}^{M \times N}$, the hidden and output vectors of LSTM cells for X are $H = [h_1, h_2, \ldots, h_N] \in \mathcal{R}^{d_h \times N}$ and $O = [o_1, o_2, \ldots, o_N] \in \mathcal{R}^{d_o \times N}$ respectively. Then the attention weight is

$$\alpha = \text{softmax}(w^T \tanh([\begin{matrix} W_h H \\ W_o O \end{matrix}])), \tag{4}$$

where $w \in \mathcal{R}^{d_h + d_o}, W_h \in \mathcal{R}^{d_h \times d_h}, W_o \in \mathcal{R}^{d_o \times d_o}$ are projection weights of attention, LSTM hidden and output cells respectively, and $\text{softmax}, \tanh$ are activation functions. Then the output of AT-LSTM module is $r = H\alpha^T \in \mathcal{R}^{d_h}$, followed by a dense layer $d = w_d^T r \in \mathcal{R}^2$ for AD classification.

Suppose the predicted output probability of the neural network is $\hat{y} = \text{softmax}(d) = f(X|\theta)$ and the expected output probability is y, where w_d are the dense layer weights, f and θ are the overall function and parameter set of the neural network respectively. Then the objective of the network is to minimize the cross entropy loss between the output probability distributions of $\hat{y}$ and y, i.e.

$$\arg\min_{\theta} - \sum_{i=0}^{1} y_i \log(f(X|\theta)_i) + \lambda \|\theta\|^2, \tag{5}$$

where λ is L-2 regularization term.

To solve Eq. 5, we update the parameters by the Adam optimizer [22], which is a stochastic optimization method with adaptive moment estimators and weight decay regularization.

3. EXPERIMENTS

3.1. Experimental setup

3.1.1. Feature Extraction

All features are derived from participant speech and corresponding transcripts from the Cookie-Theft Picture Description Task [14] of ADReSS dataset. We extracted and concatenated the participant parts in each interview by manual timestamps, including the segments of silence/filled-pauses. We then extracted the features at according to participants, which means that each participant was represented by a feature vector.

The ComParE and Linguistics feature sets are extracted by OpenSMILE [23] and CLAN [24] respectively. The TF-IDF vector is derived by term inverse document frequencies. And the X-vector and BERT embeddings are extracted by encoders that are pre-trained, as mentioned in Section 2. TF-IDF vectors and BERT embeddings are also extracted from ASR transcripts. The sizes of the ComParE and Linguistics feature sets are 6,373 and 34, and dimensions of TF-IDF, X-vector and BERT are 1,035, 512 and 768 respectively.

For feature selection, we perform Pearson's correlation test for the ComParE and Linguistics feature sets to remove features with correlation coefficients $|R| < 0.25$. Thereafter, 210 ComParE features and 10 Linguistics features are retained. Features in the ComParE set with the highest correlation are the relative spectral filtering (RASTA) representation [25], segment length, and zero cross rates group. Features in the Linguistics set that have highest correlations are verbs per utterance, mean length of an utterance and type/token ratio.

3.1.2. Implementation Details

There are mainly three systems for AD detection, i.e. Baseline LDA, the proposed SVM and AT-LSTM classifiers. In each system, we first select the features with Pearson's correlation test and reduce the feature space with PCA method, and then input these features into the classifiers.

In the pre-processing step, we perform standard normalization for all features. In the SVM system, we further perform PCA after the normalization to reduce the dimensionality. In the training step, the baseline LDA system is solved by the SVD method as Eq. (2), SVM is solved by the dual optimization method, and AT-LSTM are trained by the Adam optimizer with $5e-4$ learning rate and $1e-2$ weight decay regularization. The hidden size of LSTM cells is set to 64, with dropout rate of 0.2 to avoid over-fitting. The training epoch is 500 with 16 epochs tolerance of early stopping, and the batch size is 32.

3.1.3. Evaluation Metrics

The systems are evaluated by 10-fold cross-validation (CV) on the training data and tested on ADReSS test data. We ran 10-fold CV 10 times and averaged the resulting scores. The scores for evaluating classification performance include accuracy scores (ACC), precision (PRE), recall (REC), F1, receiver operating characteristic (ROC) curve and area under curve (AUC) with respect to the positive class (AD).

Model	Feature	ACC	PRE	REC	F1	AUC
LDA [13]	ComParE	0.56 / 0.62	0.57 / 0.60	0.52 / 0.75	0.54 / 0.67	N/A
LDA [13]	Linguistics	0.77 / 0.75	0.77 / 0.83	0.76 / 0.62	0.77 / 0.71	N/A
SVM [9]	BERT	0.82 / 0.83	0.84 / 0.81	0.79 / **0.88**	0.81 / 0.84	N/A
LDA	ComParE	0.66 / 0.65	0.65 / 0.64	0.62 / 0.62	0.64 / 0.64	0.71 / 0.66
	X-vector	0.63 / 0.58	0.62 / 0.59	0.66 / 0.54	0.62 / 0.57	0.66 / 0.63
	Linguistics	0.81 / 0.83	0.86 / 0.94	0.73 / 0.71	0.78 / 0.81	0.90 / 0.90
	TF-IDF	0.76 / 0.71	0.79 / 0.81	0.73 / 0.54	0.74 / 0.65	0.84 / 0.88
	BERT	0.76 / 0.79	0.74 / 0.79	0.80 / 0.79	0.76 / 0.79	0.83 / 0.89
SVM	ComParE	0.71 / 0.58	0.73 / 0.62	0.68 / 0.42	0.68 / 0.50	0.76 / 0.60
	X-vector	0.61 / 0.58	0.62 / 0.60	0.61 / 0.50	0.60 / 0.55	0.62 / 0.62
	Linguistics	0.80 / 0.83	0.82 / 0.90	0.75 / 0.75	0.76 / 0.82	0.89 / **0.90**
	TF-IDF	**0.86** / 0.71	**0.91** / 0.73	**0.82** / 0.67	**0.85** / 0.70	**0.93** / 0.83
	BERT	0.75 / **0.88**	0.74 / **0.91**	0.79 / 0.83	0.75 / **0.87**	0.83 / 0.89
AT-LSTM	ComParE	**0.80** / 0.64	0.81 / 0.64	0.80 / 0.64	0.79 / 0.64	0.87 / 0.71
	X-vector	0.58 / **0.67**	0.58 / 0.66	0.65 / 0.69	0.59 / 0.67	0.65 / 0.71
	Linguistics	0.82 / 0.81	0.88 / 0.88	0.76 / 0.73	0.79 / 0.79	0.90 / 0.88
	TF-IDF	0.82 / 0.66	0.84 / 0.67	0.79 / 0.65	0.80 / 0.66	0.87 / 0.77
	BERT	0.80 / 0.83	0.80 / **0.91**	0.80 / 0.74	0.78 / 0.81	0.89 / **0.90**

Table 1. Results of benchmark and proposed models on ADReSS dataset, metrics denoted as CV / Test.

3.2. Results & Analysis

3.2.1. Baseline v.s. Proposed

The classification results of LDA, SVM and AT-LSTM for different features under CV and test settings are shown in Table 1.These results show that the SVM system generally achieves the best performance over the baseline and AT-LSTM for given features. The TF-IDF and BERT features generally achieve the best performance for a given classifier. Linguistic features generally achieve better performance than acoustic features, and more robust performance when we compare performance across the CV and test settings. The best accuracy scores for acoustic features is 0.67 from the X-vector on AT-LSTM, and 0.88 from the linguistic features on BERT.

Comparing with the benchmarks in [13] shows that the improvement is significant for the ComParE and Linguistics feature sets, mostly because of the different selection and PCA methods mentioned in Section 3.

3.2.2. Manual v.s. Automatic

The classification results using TF-IDF and BERT features extracted from manual and ASR-derived transcripts are shown in Table 2, using SVM with linear kernel. The ASR systems are adopted from [21], i.e. Sys. 4 and Sys. 10, with participant word error rates (WER) 44.89% and 33.17% on ADReSS test data respectively. The results indicate that the use of ASR transcriptions for feature extraction to detect AD achieves results that are as good as the use of manual ones. The best performance come from BERT features extracted from ASR Sys. 10, with same accuracy and better AUC compared with manual one.

System	Feature	ACC	PRE	REC	F1	AUC
Sys. 4 (0.45)	TF-IDF	0.69	**0.74**	0.58	0.65	0.85
	BERT	**0.79**	0.72	**0.96**	**0.82**	**0.87**
Sys. 10 (0.33)	TF-IDF	0.69	0.74	0.58	0.65	0.82
	BERT	**0.88**	**0.82**	**0.96**	**0.88**	**0.92**
Manual	TF-IDF	0.71	0.73	0.67	0.70	0.83
	BERT	**0.88**	**0.91**	**0.83**	**0.87**	**0.89**

Table 2. Test results of manual and ASR-based features. (*) denotes WER.

4. DISCUSSION & CONCLUSION

In this work, we have presented a comparative study of different speech and language features for AD detection using different classifiers. The features we explored include the ComParE set, X-vectors, the Linguistics set, TF-IDF and BERT embeddings, extracted from speech, manual or ASR transcripts. The models we proposed include LDA, SVM and AT-LSTM, covering unsupervised, supervised and transfer learning.

The size of the dataset limits the performance of neural networks, and we find that Pearson's correlation test and PCA are useful for improving the classification performance. We also find that the features with highest correlation with the presence/absence of AD include acoustic features (namely, the RASTA-style filtered auditory spectrum, segment length and zero cross rate and their statistical functionals), and linguistic features (namely, verbs, mean lengths of utterances and type/token ratios).

Experiments on the ADReSS dataset indicate that the use of acoustic and linguistic features are viable for AD detection, and linguistic features outperform acoustic features. Results also indicate that linguistic features extracted from ASR transcripts can achieve detection performance as good as manual transcripts. This result suggests the feasibility of fully automatic AD detection from speech and language features.

Future work will include applying the proposed approach to native Cantonese data, as well as feature fusion to boost performance on AD detection.

5. ACKNOWLEDGEMENTS

This project is partially supported by the HKSARG Research Grants Council's Theme-based Research Grant Scheme (Project No. T45-407/19N).

6. REFERENCES

[1] P. S. Sachdev, D. Blacker, D. G. Blazer, et al., "Classifying neurocognitive disorders: the DSM-5 approach," *Nature Reviews Neurology*, vol. 10, no. 11, pp. 634, 2014.

[2] Alzheimer's Disease International, "World Alzheimer report 2019: attitudes to dementia," 2019.

[3] L. Velayudhan, S.-H. Ryu, M. Raczek, et al., "Review of brief cognitive tests for patients with suspected dementia," *International psychogeriatrics*, vol. 26, no. 8, pp. 1247–1262, 2014.

[4] G. Szatloczki, I. Hoffmann, V. Vincze, et al., "Speaking in Alzheimer's disease, is that an early sign? Importance of changes in language abilities in Alzheimer's disease," *Frontiers in aging neuroscience*, vol. 7, pp. 195, 2015.

[5] 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.

[6] J. Weiner, C. Frankenberg, J. Schröder, and T. Schultz, "Speech reveals future risk of developing dementia: Predictive dementia screening from biographic interviews," in *2019 IEEE ASRU*. IEEE, 2019, pp. 674–681.

[7] M. L. B. Pulido, J. B. A. Hernández, M. Á. F. Ballester, et al., "Alzheimer's disease and automatic speech analysis: A review," *Expert Systems with Applications*, vol. 150, pp. 113213, 2020.

[8] J. V. E. López, L. Tóth, I. Hoffmann, et al., "Assessing Alzheimer's Disease from speech using the i-vector approach," in *International Conference on Speech and Computer*. Springer, 2019, pp. 289–298.

[9] A. Balagopalan, B. Eyre, F. Rudzicz, and J. Novikova, "To BERT or not to BERT: Comparing speech and language-based approaches for Alzheimer's Disease detection," *Proc. Interspeech 2020*, pp. 2167–2171, 2020.

[10] J. Yuan, Y. Bian, X. Cai, et al., "Disfluencies and fine-tuning pre-trained language models for detection of alzheimer's disease," *Proc. Interspeech 2020*, pp. 2162–2166, 2020.

[11] Y. Wang, M. Huang, X. Zhu, and L. Zhao, "Attention-based LSTM for aspect-level sentiment classification," in *Proceedings of the 2016 conference on empirical methods in natural language processing*, 2016, pp. 606–615.

[12] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[13] S. Luz, F. Haider, S. de la Fuente, et al., "Alzheimer's Dementia Recognition Through Spontaneous Speech: The ADReSS Challenge," *Proc. Interspeech 2020*, pp. 2172–2176, 2020.

[14] J. T. Becker, F. Boiler, O. L. Lopez, 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, 1994.

[15] A. Satt, R. Hoory, A. König, et al., "Speech-based automatic and robust detection of very early dementia," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[16] F. Weninger, F. Eyben, B. W. Schuller, et al., "On the acoustics of emotion in audio: what speech, music, and sound have in common," *Frontiers in psychology*, vol. 4, pp. 292, 2013.

[17] D. Snyder, D. Garcia-Romero, G. Sell, et al., "X-vectors: Robust DNN embeddings for speaker recognition," in *2018 IEEE ICASSP*. IEEE, 2018, pp. 5329–5333.

[18] J. Ramos et al., "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*. New Jersey, USA, 2003, vol. 242, pp. 133–142.

[19] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," *Proc. Interspeech 2017*, pp. 999–1003, 2017.

[20] 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.

[21] Z. Ye, S. Hu, J. Li, et al., "Development of the CUHK elderly speech recognition system for neurocognitive disorder detection using the DementiaBank corpus," in *submission to ICASSP*. IEEE, 2021.

[22] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[23] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[24] B. MacWhinney and J. Wagner, "Transcribing, searching and data sharing: The CLAN software and the TalkBank data repository," *Gesprachsforschung: Online-Zeitschrift zur verbalen Interaktion*, vol. 11, pp. 154, 2010.

[25] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE transactions on speech and audio processing*, vol. 2, no. 4, pp. 578–589, 1994.