INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Automatic Detection of Alzheimer's Disease Using Spontaneous Speech Only

Jun Chen[1], *Jieping Ye*[1], *Fengyi Tang*[2], *Jiayu Zhou*[2]

[1]Department of Computational Medicine and Bioinformatics, University of Michigan, USA
[2]Department of Computer Science and Engineering, Michigan State University, USA
junnchen, jpye@umich.edu, jiayuz, tangfeng@msu.edu

Abstract

Alzheimer's disease (AD) is a neurodegenerative syndrome which affects tens of millions of elders worldwide. Although there is no treatment currently available, early recognition can improve the lives of people with AD and their caretakers and families. To find a cost-effective and easy-to-use method for dementia detection and address the dementia classification task of InterSpeech 2021 ADReSSo (Alzheimer's' Dementia Recognition through Spontaneous Speech only) challenge, we conduct a systematic comparison of approaches to detection of cognitive impairment based on spontaneous speech. We investigated the characteristics of acoustic modality and linguistic modality directly based on the audio recordings of narrative speech, and explored a variety of modality fusion strategies. With an ensemble over top-10 classifiers on the training set, we achieved an accuracy of 81.69% compared to the baseline of 78.87% on the test set. The results suggest that although transcription errors will be introduced through automatic speech recognition, integrating textual information generally improves classification performance. Besides, ensemble methods can boost both the accuracy and the robustness of models.

Index Terms: Cognitive Decline Detection, Modality Fusion, Alzheimer's Disease, Computational Paralinguistics

1. Introduction

Dementia is a syndrome associated with a deterioration in memory, language, problem-solving, and other cognitive functions to perform daily activities. According to the estimation of WHO, there are around 50 million people having dementia worldwide, and this number is increasing by nearly 10 million every year. Amongst many different forms of dementia, Alzheimer's disease is the most common one and contributes to 60–70% of total cases [1]. Screening of Alzheimer's dementia is typically conducted through paper-and-pencil cognitive tests, such as the Mini Mental Status Examination (MMSE) [2] and the Montreal Cognitive Assessment (MoCA) [3]. Although cheap and quick to administer, the scoring process totally relies on the personal judgment of clinicians, which may introduce errors and result in a high inter-rater variability [4]. To address these issues, extensive studies have been carried out for the purpose of automated cognitive assessment [5]. One promising direction is speech-based screening. Speech signals can be relatively easily collected throughout the day without burdening the participants or the researchers. Moreover, the rapid development of speech technology and machine learning algorithms provides us a good opportunity to utilize those speech data for automatic screening of dementia [6] and finally translate speech-based methods into clinical practice.

There are existing efforts on the acoustic characteristics of AD. In [7], Warnita et al. extracted several sets of paralinguistic features from speech utterances in DementiaBank Pitt Corpus [8]. They trained a gated convolutional neural network for utterance-level AD classification, and then made the final verdict for each subject through majority voting. The best accuracy of 73.6% was achieved from this acoustic-only method. Luz et al. [9] assessed the effectiveness of several other acoustic feature sets with different classifiers for AD detection on the same data set. They showed that the eGeMAPS feature set provided the best single feature set accuracy and simple hard fusion of feature sets could improve the accuracy from 71.34% to 78.70%. [10] used low-level descriptors of IS10-Paralinguistics feature set [11] and Bag-of-Acoustic-Words (BoAW) for feature aggregation, and achieved a leave-one-subject-out (LOSO) accuracy of 76.85% on the training set of InterSpeech 2020 ADReSS challenge. Moreover, there are evidences showing that using manual transcripts of speech or a combination of transcripts and speech audio can generally lead to better performance compared to using audio alone. In [12], Yuan et al. explored disfluencies and language problems in Alzheimer's Disease subjects, and achieved the best accuracy 89.6% on the test set of the ADReSS challenge by fine-tuning Transformer-based pre-trained language models. Syed et al. achieved an accuracy of 85.45% on the same challenge task [10] by using both acoustic features and linguistic features from manual transcription.

In this paper, we conducted a systematic comparison of methods of detecting cognitive impairment based on a narrative speech from a picture description task, and extensively explored different modality fusion strategies. We first investigated the characteristics of acoustic modality and trained machine learning classifiers based on speech paralinguistic feature sets. Next, we generated two sets of transcripts through automatic speech recognition (ASR) and extracted linguistic features, both deep text embedding, and human-defined psychological features, from transcripts for dementia screening. Finally, we compared different modality fusion strategies to boost the performance of our model. Our proposed model outperformed the ADReSSo challenge baseline for AD classification task on both training partition and test partition.

2. Dataset

The ADReSSo challenge [6] released two distinct benchmark datasets for three different tasks. The dataset for AD classification consists of audio recordings of a picture descriptions task from both cognitively healthy people and patients diagnosed with AD. Participants were asked to describe the *Cookie Theft* picture from the Boston Diagnostic Aphasia Examination [8]. Recordings were preprocessed with stationary noise removal and audio volume normalization across audio segments to reduce the variation caused by recording conditions [6].

The resulting dataset includes 237 audio files. To minimize the risk of bias in the prediction, these files are carefully partitioned into training and test sets at a ratio of 7:3 so as to preserve

http://dx.doi.org/10.21437/Interspeech.2021-2002

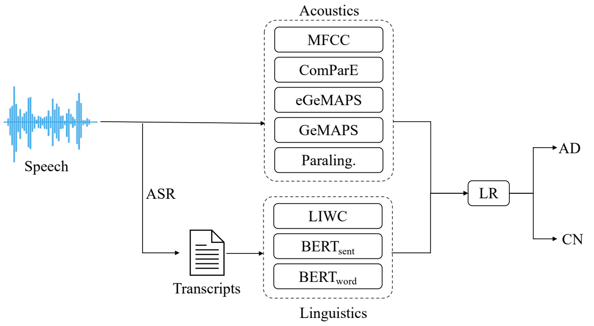

Figure 1: *Framework of automatic screening of AD.*

the balance of gender and age distribution [6]. There are 166 instances allocated to the training set. Among them, 87 subjects are diagnosed with dementia and 79 are elderly normal controls. The other 71 instances are allocated to the test set, among which 35 are with an AD diagnosis and 36 are cognitively normal.

3. Methodology

In this paper, we investigate the efficacy of logistic regression (LR) model on AD screening, as well as modality fusion strategies to boost the performance of the model. The schematic workflow of audio-based and ASR transcripts-based screening is shown in Figure 1.

3.1. Screening based on speech

In this section, we use python library librosa [13] and open-source audio feature extraction toolkit openSMILE [14] for audio preprocessing and paralinguistic acoustic feature extraction. Paralinguistics have been widely used for emotion recognition and detection of some other mental disorders such as depression [15] and bipolar disorder [16]. Evidence shows that AD patients have deterioration in emotional control [1] and may have difficulty in expressing emotions in prosodies [17]. Based on this, we hypothesize that paralinguistic acoustics is a good candidate for AD biomarkers. Five sets of acoustic features which are known to represent paralinguistic characteristics of speech are extracted as follows.

1. Mel-frequency cepstral coefficients (MFCCs)[18]

   The first 13 MFCC bands (0-12), and corresponding 13 delta MFCCs and 13 delta-delta MFCCs, which reflect the rate of change and the acceleration in MFCCs, are extracted. Descriptive statistics functions are applied, totaling 468 features for one utterance.
2. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS)[19]

   GeMAPS contains the 18 low-level descriptors(LLD), including frequency related parameters pitch, jitter, formants, energy related parameters shimmer, loudness, harmonics-to-noise ratio (HNR), and several spectral parameters. Statistical functionals are applied to each LLD, totalling 62 parameters.
3. The extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS)[19]

   An extension set, which contains 7 LLDs of cepstral and dynamic parameters and corresponding functionals, is added to the GeMAPS. In total, 88 features are extracted per utterance.
4. INTERSPEECH 2016 Computational Paralinguistics Challenge Feature Set (ComParE-2016)[20]

   ComParE is the largest standard set of openSMILE with 6373 features. It is a brute-force feature set that has proved to be useful for a variety of speech paralinguistic tasks. ComParE-2016 is the most recent version of the ComParE.
5. INTERSPEECH 2010 Paralinguistics Challenge Feature Set (IS10-Paraling.)[11]

   IS10-Paralinguistics contains 38 LLDs and 38 corresponding delta coefficients. 21 functionals are applied to these LLDs, totaling 76 LLD for one frame and 1582 features for one utterance. This feature set can be considered as a low-dimensional alternate to the ComParE.

Extracted feature sets are normalized by standard min-max scaling and passed down to the logistic regression classifier as illustrated in the machine learning pipeline in Figure 1.

3.2. Screening based on auto-transcription

Language impairment is a distinguishing marker of dementia [5]. In addition to analyzing acoustic characteristics of AD, we also utilize ASR techniques to introduce the second modality, which is transcript text. Three sets of linguistic features are then extracted from each auto-transcript, as illustrated in Figure 1.

3.2.1. Automated transcript generation through ASR

We use two different approaches to generate transcripts from spontaneous speech. The first approach is using a pre-trained English DeepSpeech model from Mozilla [21]. The second is using Google Cloud standard speech-to-text service. Both transcribing processes can be done automatically, with no need for model fine-tuning or human intervention. Linguistic feature sets are extracted from transcripts produce by each approach.

3.2.2. Linguistic feature based screening

The language feature sets can be largely categorized into two classes. One is a transparent linguistic feature set, in which each individual feature has a well-defined meaning and thus having desirable interpretability. Linguistic Inquiry and Word Count (LIWC) [22] is one such text analysis method. It counts words in psychologically meaningful categories. Extensive studies using LIWC demonstrated its ability to detect meaning in a wide variety of experimental settings, such as attentional focus, emotion, and thinking styles. Here we extract a 64-dimensional LIWC vector from each transcript.

Another category of linguistic feature is language embedding which usually a dense numerical representation of a word or a sentence. Those embeddings have already shown great success in a variety of tasks of natural language processing [23]. We investigate the efficacy of pre-trained embeddings from deep language model Bidirectional Encoder Representations from Transformers (BERT) [24]. Specifically, we compute embedding for each word(BERT$_{\text{word}}$) and embedding for the whole transcript(BERT$_{\text{sent}}$) from pretrained BERT base uncased model using the Huggingface Transformers library[25]. BERT$_{\text{sent}}$, as a 768-dim linguistic feature, is passed to the downstream pipeline directly, while BERT$_{\text{word}}$ are first aggregated by four kinds of pooling functions (max, min, average, and standard deviation) to generate a transcript-level representation. The outputs of

Table 1: *Summary of results for AD classification using acoustic features on LOSO cross-validation*

Feature Set	Accu.	F1	Spec.
MFCC	67.47	69.32	64.56
GeMAPS	68.67	70.79	64.56
eGeMAPS	**74.10**	**75.98**	69.62
ComParE	71.69	72.83	70.89
Paraling.	71.69	72.19	**73.42**
Maj. voting	**77.11**	**78.41**	**74.68**
Avg. fusion	74.7	75.58	**74.68**
Wgt. fusion	74.7	75.58	**74.68**

pooling operations are concatenated, resulting in a 6373-dim vector, and passed down to the LR classifier (Figure 1).

3.2.3. Modality fusion

To make full use of audio and text modalities, we explored a multimodal framework for the automatic screening of AD. We first apply a straightforward early fusion or data-level fusion [26] by selecting a good performing feature set in each modality, and concatenating features into a single vector for each sample. Different combinations of feature sets are investigated for comparison. Besides, we also applied the late fusion (decision level fusion) strategy which uses input modalities independently followed by fusion at a decision-making stage. It is inspired by the popularity of ensemble methods [27]. Three different rules are used in this paper to combine independently trained classifiers: a) majority voting of predicted class labels, b) average fusion of predicted class probabilities, and c) weighted average fusion of class probabilities with the weight set as the accuracy of the corresponding base classifiers.

4. Experiments and Results

We use LR classifier for the AD classification task. Two hyperparameters are fine-tuned for optimized performance. Regularization strength λ was tuned with grid search between a range of $5e$-5 and $1e2$, and penalty is chosen from $\{$L1, L2$\}$.

4.1. LOSO evaluation of acoustic models

In Table 1, we show the classification results of five acoustic feature sets. eGeMAPS outperforms other sets with a leave-one-subject-out (LOSO) accuracy of 74.10% on the training set. This result is closely followed by the Paraling. and ComParE models which achieve the second-best single model accuracy of 71.69%. Since Paraling. is a low-dimensional alternate to the ComParE set, these tied results are as expected. The model based on GeMAPS achieved an accuracy of 68.67%, which is 5.43% lower than the eGeMAPS based model. This indicates that the 26 additional features in the extended minimalistic set do have some contribution to dementia recognition. Furthermore, the ensemble methods can boost the overall performance of audio modality to 77.11%.

4.2. LOSO evaluation of linguistic models

From the summary in Table 2, we see that models based on transcripts generated by Google Cloud universally outperform models based on DeepSpeech transcripts. Among them, BERT word embedding model achieved the best LOSO accuracy of 75.30%. This result is followed by the result of another embedding based model BERT$_{\text{sent}}$. LIWC based models from both transcription settings have an accuracy of less than 70%.

Table 2: *Summary of results for AD classification using linguistic features on LOSO cross-validation*

ASR Model	Feature Set	Accu.	F1	Spec.
DeepSpeech	LIWC	67.47	71.28	56.96
	BERT$_{\text{word}}$	68.07	71.04	60.76
	BERT$_{\text{sent}}$	70.48	73.22	63.29
	Maj. Voting	69.88	72.53	63.29
	Avg. Fusion	66.87	69.95	59.49
	Wgt. Fusion	67.47	70.33	60.76
GoogleCloud	LIWC	69.88	71.26	68.35
	BERT$_{\text{word}}$	**75.30**	**76.30**	**74.68**
	BERT$_{\text{sent}}$	72.89	74.58	69.62
	Maj. Voting	75.30	76.84	72.15
	Avg. Fusion	71.08	72.41	69.62
	Wgt. Fusion	71.08	72.41	69.62
Overall	Maj. Voting	73.49	75.0	70.89
	Avg. Fusion	72.29	74.44	67.09
	Wgt. Fusion	72.29	74.44	67.09

We noticed that the average number of words is 60.9 for transcripts generated by pretrained DeepSpeech and 91.2 for those generated by Google Cloud speech-to-text service, which indicates that the latter may have a better speech recognition performance on this particular data set. This partially explains why models based on the latter generally have better classification performance.

4.3. Modality Fusion

An ideal feature set should be compact enough to be implemented in a real-time system and robust enough to detect subtle changes of spontaneous speech of people developing dementia. Here we used early and late fusion strategies to investigate the multi-modality classification problem and the results are summarized in Table 3. Three linguistic feature sets based on Google Cloud transcripts are combined with each one of the top three acoustic feature sets in Section 4.1. The best accuracy is achieved by simple early fusion of google-LIWC (g-LIWC) and eGeMAPS. This combination also produces the most compact feature set (152 features in total) in our modality fusion settings. Furthermore, feature sets from both modalities are interpretable, which can help us to have a better understanding of the linguistic and paralinguistic characteristics of the disease. The second-best performance comes from the late average fusion of google-BERT$_{\text{word}}$ (g-BERT$_{\text{word}}$) and eGeMAPS, which is higher than the early fusion of the same sets. Late fusion, however, does not always outperform early fusion. Late fusion of g-LIWC and eGeMAPS, for example, has an accuracy 9% lower than the early fusion. Besides, we also applied ensemble methods to combine the predictions of either all of the twenty single classifiers or the selected top ten classifiers. An early fusion model is considered a single classifier here since it is trained only once after feature concatenation. The best ensemble accuracy is 80.72% following the majority voting rule.

Table 3: *Results of multimodal methods on training set LOSO cross-validation*

FS	Feature Set	Accu.	F1	Spec.
Early	g-LIWC+eGeMAPS	**81.93**	**82.56**	**82.28**
	g-BERT$_{\text{word}}$+eGeMAPS	75.90	76.74	75.95
	g-BERT$_{\text{sent}}$+eGeMAPS	74.10	75.71	70.89
	g-LIWC+ComParE	71.69	73.45	68.35
	g-BERT$_{\text{word}}$+ComParE	74.70	75.58	74.68
	g-BERT$_{\text{sent}}$+ComParE	75.30	76.02	75.95
	g-LIWC+Paraling.	75.90	76.74	75.95
	g-BERT$_{\text{word}}$+Paraling.	77.11	77.91	77.22
	g-BERT$_{\text{sent}}$+Paraling.	78.92	80.23	75.95
Late	g-LIWC + eGeMAPS	72.29	74.16	68.35
	g-BERT$_{\text{word}}$ + eGeMAPS	**77.71**	**78.61**	**77.22**
	g-BERT$_{\text{sent}}$ + eGeMAPS	75.90	77.01	74.68
	g-LIWC + ComParE	71.08	72.09	70.89
	g-BERT$_{\text{word}}$ + ComParE	72.89	74.29	70.89
	g-BERT$_{\text{sent}}$ + ComParE	72.29	73.26	72.16
	g-LIWC + Paraling.	73.49	74.42	73.42
	g-BERT$_{\text{word}}$ + Paraling.	72.29	74.73	65.82
	g-BERT$_{\text{sent}}$ + Paraling.	71.08	71.76	72.15
Ens.	Overall Avg.	75.3	76.84	72.15
	Overall Maj.	77.11	79.12	70.89
	Overall Wgt.	80.12	80.92	79.75
	Top-10 Avg.	79.52	80.23	79.75
	Top-10 Maj.	**80.72**	**81.18**	**82.28**
	Top-10 Wgt.	79.52	80.23	79.75

4.4. Predictions for the test partition

ADReSSo challenge allows each team to submit the results of five attempts. A summary of the baseline and our results for the test set is provided in Table 4. For the first attempt, we use predictions from the early fusion of g-LIWC and eGeMAPS, which was the best performing model on the training partition by achieving an accuracy of 81.93%. On the test set, however, this model only achieved an accuracy of 64.79%, which indicates overfitting on the training set. The second attempt is the early fusion of g-BERT$_{\text{sent}}$ and Paraling., which is the second-best early fusion model. The performance dropped a little bit on the test set from 77.11% to 74.65%. The third attempt is the majority voting of all the classifiers from audio modality which achieved an accuracy of 67.61% on the test set. This indicates that the information that audio modality offers is not robust and sufficient enough for AD detection. The fourth and fifth attempts used ensemble strategies. Majority voting and average fusion are applied separately on the top 10 best-performing models. The resultant prediction accuracy scores for the test partition are 80.28% and 81.69%, which are both better than the challenge baseline of 78.87%. We also noticed that these two models achieved similar accuracy on the training set and test set, which shows that the ensemble increased the robustness of the models.

Table 4: *Summary of results for AD classification on test set*

Feature Set	Accu.	F1	Spec.
g-LIWC+eGeMAPS	64.79	72.22	61.54
g-BERT$_{\text{sent}}$+Paraling.	74.65	80.56	72.73
Audio Maj.	67.61	77.78	63.49
Top-10 Maj.	80.28	88.89	78.13
Top-10 Avg.	**81.69**	**88.89**	**80.00**
Challenge baseline[6]	78.87	77.78	78.87

5. Discussion

The effectiveness of several paralinguistic feature sets for Alzheimer's recognition was evaluated in Section 4.1. In addition to utilizing those predefined features sets, one future direction could be introducing paralinguistic embeddings generated from a representation model pretrained on a large external dataset to our dementia recognition pipeline. While representation learning models like BERT have achieved great success in the text domain, such methods are underutilized in the speech domain.

Our model built on linguistic and acoustic features achieved the best accuracy of 81.69% on the test set, while paralinguistics based model only had an accuracy of 67.61%. This indicates that although paralinguistic changes are potential markers of dementia, acoustic modality alone may not have enough information for the diagnosis of disease. Introducing linguistic features through ASR often leads to a considerable improvement to the predictions accuracy of the disease, despite the fact that the ASR transcripts have a relatively high word error rate.

Another observation is that, from the aspect of accuracy and robustness, ensemble methods generally gives better performance, especially when we have lots of individual classifiers. This is because the errors from multiple models are dealt with independently.

We also noticed that the most promising model based on eGeMAPS and g-LIWC on the training set performs worst among the five attempts on the test partition. One explanation for the performance gap between training and test stages might be the disadvantage of LOSO cross validation. Even though test-error is unbiased in each iteration, LOSO has a high variability as only one observation is predicted for validation. Stratified 10-fold cross-validation or nested cross validation could be applied in future study to alleviate the overfitting issue caused by LOSO. Another possible explanation might be the choice of base classifier. Our logistic regression model marginally outperformed other machine learning classifiers, including SVM, decision tree, and multilayer perceptron, regarding the LOSO performance on the training set, while other classifiers might be more robust to the outliers or have a better generalization capability.

6. Acknowledgement

This material is based in part upon work supported by the National Science Foundation under Grant IIS-1749940, Office of Naval Research N00014-20-1-2382, and National Institute on Aging (NIA) RF1AG072449.

7. References

[1] “Dementia,” Sep 2020. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dementia

[2] M. F. Folstein, S. E. Folstein, and P. R. McHugh, ““mini-mental state”: a practical method for grading the cognitive state of pa-

tients for the clinician," *Journal of psychiatric research*, vol. 12, no. 3, pp. 189–198, 1975.

[3] Z. S. Nasreddine, N. A. Phillips, V. Bédirian, S. Charbonneau, V. Whitehead, I. Collin, J. L. Cummings, and H. Chertkow, "The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment," *Journal of the American Geriatrics Society*, vol. 53, no. 4, pp. 695–699, 2005.

[4] S. Chen, D. Stromer, H. A. Alabdalrahim, S. Schwab, M. Weih, and A. Maier, "Automatic dementia screening and scoring by applying deep learning on clock-drawing tests," *Scientific Reports*, vol. 10, no. 1, pp. 1–11, 2020.

[5] L. Tóth, I. Hoffmann, G. Gosztolya, V. Vincze, G. Szatlóczki, Z. Bánréti, M. Pákáski, and J. Kálmán, "A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech," *Current Alzheimer Research*, vol. 15, no. 2, pp. 130–138, 2018.

[6] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Detecting cognitive decline using speech only: The adresso challenge," *medRxiv*, 2021.

[7] T. Warnita, N. Inoue, and K. Shinoda, "Detecting alzheimer's disease using gated convolutional neural network from audio data," *arXiv preprint arXiv:1803.11344*, 2018.

[8] J. T. Becker, "The natural history of alzheimer's disease," *JAMA Neurology*, vol. 51, no. 6, p. 585, 1994.

[9] F. Haider, S. De La Fuente, and S. Luz, "An assessment of paralinguistic acoustic features for detection of alzheimer's dementia in spontaneous speech," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 272–281, 2019.

[10] M. S. S. Syed, Z. S. Syed, M. Lech, and E. Pirogova, "Automated screening for alzheimer's dementia through spontaneous speech," *INTERSPEECH (to appear)*, pp. 1–5, 2020.

[11] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. S. Narayanan, "The interspeech 2010 paralinguistic challenge," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[12] J. Yuan, Y. Bian, X. Cai, J. Huang, Z. Ye, and K. Church, "Disfluencies and fine-tuning pre-trained language models for detection of alzheimer's disease," *Proc. Interspeech 2020*, pp. 2162–2166, 2020.

[13] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8. Citeseer, 2015, pp. 18–25.

[14] 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.

[15] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015.

[16] Z. S. Syed, K. Sidorov, and D. Marshall, "Automated screening for bipolar disorder from audio/visual modalities," in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, 2018, pp. 39–45.

[17] G. Tosto, M. Gasparini, G. Lenzi, and G. Bruno, "Prosodic impairment in alzheimer's disease: assessment and clinical relevance," *The Journal of neuropsychiatry and clinical neurosciences*, vol. 23, no. 2, pp. E21–E23, 2011.

[18] S. Molau, M. Pitz, R. Schluter, and H. Ney, "Computing mel-frequency cepstral coefficients on the power spectrum," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (cat. No. 01CH37221)*, vol. 1. IEEE, 2001, pp. 73–76.

[19] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *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.

[20] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, K. Evanini *et al.*, "The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language," in *17TH Annual Conference of the International Speech Communication Association (Interspeech 2016), Vols 1-5*, 2016, pp. 2001–2005.

[21] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[22] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: Liwc 2001," *Mahway: Lawrence Erlbaum Associates*, vol. 71, no. 2001, p. 2001, 2001.

[23] S. Ghannay, B. Favre, Y. Esteve, and N. Camelin, "Word embedding evaluation and combination," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016, pp. 300–305.

[24] 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.

[25] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Huggingface's transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019.

[26] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, "Multisensor data fusion: A review of the state-of-the-art," *Information fusion*, vol. 14, no. 1, pp. 28–44, 2013.

[27] L. I. Kuncheva, *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons, 2014.