Static vs. Dynamic Modelling of Acoustic Speech Features for Detection of Dementia

Muhammad Shehram Shah Syed[1]
RMIT University, Australia

Zafi Sherhan Syed[2]
Mehran University, Pakistan

Elena Pirogova[3]
RMIT University, Australia

Margaret Lech[4]
RMIT University, Australia

Abstract—**Dementia is a chronic neurological disease that causes cognitive disabilities and significantly impacts daily activities of affected individuals. It is known that early detection of dementia can improve the quality of life of patients through a specialized care program. Recently, there has been a growing interest in speech-based screening of neurological diseases such as dementia. The focus is on continuous monitoring of changes in speech of dementia patients, aiming to identify the early onset of the disease which could facilitate development of preventative treatment care. In this work, we propose a dynamic (temporal) modeling of acoustic speech characteristics aiming at identifying the signs of dementia. The classification performance of the proposed framework is compared with a baseline static modeling of acoustic speech features. Experimental results show that the proposed dynamic approach outperforms the static method. It achieves the classification accuracy of 74.55% compared to 66.92% obtained using the static models.**

Keywords*—*Dementia detection; speech classification; neural networks; recurrent neural networks

I. Introduction

Dementia affects a large number of people worldwide. It is estimated that currently more than 50 million individuals suffer from this disease and the number is expected to grow to 75.63 million by the end of 2030 [1].Dementia is an umbrella term for a set of progressive neurological diseases that lead to impairment or even complete loss of language, memory, thought processes, and problem-solving abilities which compromise the quality of living in affected individuals. There are various types of dementia. In Alzheimer's dementia, nerve cells connections and communication are impaired, eventually leading to nerve cells death and tissue loss throughout the brain. Over time, the brain shrinks dramatically, affecting nearly all its functions. Dementia can also be caused by prolonged suffering from high blood pressure as well as strokes, known as vascular dementia) [2], [3]. Alzheimer's disease disrupts both the way electrical charges travel within nerve cells and the activity of neurotransmitters, thus affecting functions of memory, movement, and thinking ability, which depend on the region of the brain being affected.

Traditional methods for diagnosis are based on neurophysiological tests [4], [5] and neuroimaging (MRI) [6]. However, in recent years there has been a growing focus on less invasive sensing technologies, in particular, speech-based diagnosis and monitoring of dementia [7]. This year, at the Interspeech 2020 conference, Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) Challenge was organized which encouraged researchers to develop automated methods for detecting Alzheimer's dementia from speech recordings of patients [8].

This study is focused on developing a method for automated detection of dementia using the dataset provided as part of the ADReSS challenge. Here, we propose a framework based on temporal modelling of acoustic features and demonstrate its effectiveness for the task of identifying individuals with dementia. The performance of the temporal models is bench-marked against the static models which are based on functionals of descriptive statistics. The paper is organized as follows: In section II, we present a brief literature review, in section III we provide a summary of the ADReSS challenge dataset. In section IV we detail the methodology of the temporal modelling framework. Experimental results and discussion are provided in section V and section VI, summary and future outlook are presented in section VII.

II. Related Work

The ADReSS dataset, central to this study, is the baseline paper relevant to the ADReSS challenge by Luz et al. [8]. The dataset consists of speech recordings of subjects from two groups, healthy individuals and Alzheimer's dementia sufferers. For the classification baseline, Luz et al. computed four types of acoustic feature sets, (i) emobase [9], (ii) Computational Paralinguistics Challenge (ComParE) [10], (iii) Extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [11], and (iv) Multi-Resolution Cochleagram (MRCG) [12], to represent speech characteristics of subjects from the two groups. Here, the functionals based static modelling was used to generate a representation for speech recordings using the above mentioned acoustic feature sets. Results showed that ComParE provided the best classification performance amongst the four feature sets. In [13], Haider et al. investigated the feasibility of paralinguistic features for recognizing Alzheimer's dementia from recordings of spontaneous speech. Notably, while the authors used the same feature sets as [8], they found the eGeMAPS feature set to provide the best classification performance. This suggests that the efficacy of feature sets largely depends on the dataset itself.

Literature shows that it has been a common approach of using linguistic features for the task of recognizing dementia from speech. Fraser et al. [14] computed a large number of features to capture linguistic phenomena, such as grammar constituents, information content, part-of-speech, psycholinguistics, the richness of vocabulary, and the syntactic complexity of speech transcripts. In addition to these features, Fraser et al. also investigated the existence of acoustic abnormality using features derived from the Mel Frequency Cepstral Coefficients (MFCCs) [15] which provide information about the spectral characteristics of speech. Based on their investigation, it was reported that individuals with dementia have a semantically impoverished, syntactically and information deficient language, in addition to abnormal speech. Another notable research by Mirheidari et al. [16], the authors constructed a feature vector consisting of acoustic and linguistic features for the task at hand. In addition, they also computed conversational features that are tuned towards identifying individuals with memory disorders [17]. Their research findings suggest that conversational features provide the best classification performance when transcripts are annotated manually. However, when automated speech recognition was used, the classification accuracy significantly decreased from 96.70% to 76.70%. A decrease in accuracy was also observed for linguistic features with automatically generated transcripts. These results highlight the limitations of automated screening methods based on transcripts, especially when high-fidelity speech transcription is not possible, which can be the case for people suffering from diseases affecting their speaking capability. In the current work, therefore, we focus only on the audio modality.

III. Dataset

To validate the proposed methodology, we used the dataset provided by the ADReSS challenge at Interspeech 2020. This is advantageous as our experiments can be reproduced by other researchers, since the dataset is available in the public domain. The dataset includes speech recordings from 144 subjects in total; one-half of those are individuals with dementia, whereas the other half are healthy individuals. The recordings have an average duration of 75.30 seconds, a standard deviation of 38.38 seconds, and a maximum duration of 268.48 seconds. Given the significant variation in the duration of speech recordings, we segment each recording into 10 seconds based on non-overlapping chunks for training classifiers on equal duration of speech. At the evaluation stage of the classifier, majority voting was conducted for classification performance analysis. This means that the class with higher probability was assigned to the input sample. A summary of dataset distribution is provided in Table I.

TABLE I. Distribution of Subjects with Dementia and those who are Healthy in the ADReSS Dataset

Gender	Label	
	Healthy	**Dementia**
Male	36	36
Female	42	42
Σ	72	72

IV. Methodology

We hypothesize that temporal characteristics of speech acoustics can be useful for distinguishing between healthy and dementia individuals. The hypothesis is based on the fact that patients with dementia reveal a lack of speech fluency and exhibit other rhythmic issues (pause and forget, difficulty joining or following a conversation).

For this analysis, we started by computing low-level acoustic speech descriptors (LLDs) preserving the paralinguistic aspect of speech. The LLDs are extracted from speech waveforms segment by segment, thus preserving the temporal characteristics. We normalize the LLD features using standard scaling (z-scores), and then pass them to a recurrent sequence network for generating an embedding that preserves the temporal characteristics of speech. Finally, a fully connected dense classifier is applied to identify subjects with dementia and healthy. The functional block diagram of the proposed temporal modelling framework is shown in Fig. 1.

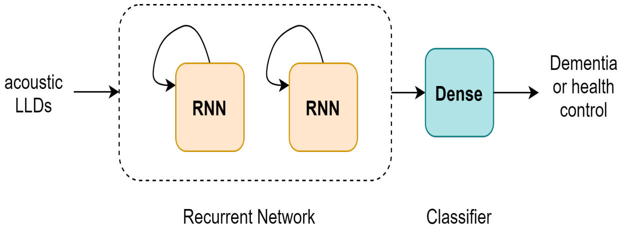

Fig. 1. Block Diagram of Temporal (Dynamic) Modelling of Speech Acoustic Features.

A. Audio Features for Speech Paralinguistics

The term speech paralinguistics refers to non-verbal and non-linguistic aspects of speech. As per Schuller et al. [18], paralinguistics are important facets of communication, when human-beings naturally communicate their underlying emotional states without explicitly describing them. It has been shown that audio (acoustic) features representing speech paralinguistics can also be used to screen individuals for various disorders, such as autism, bipolar disorder, depression, dementia, and Parkinson's disease (they can effectively characterize manifestations of mental and neurological disorders based on speech) [19], [20], [21], [13]. Here, we utilize three expert-knowledge based acoustic feature sets that are known to adequately represent characteristics of speech paralinguistics. These feature sets include the Interspeech 2010 Paralinguistics Challenge feature set (IS10-Paralinguistics) [22], the Interspeech 2013 Computational Paralinguistics Challenge (ComParE) feature set [18], and the Extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) feature [11]. Given that these feature sets are well known, we refer the reader to [22], [18], [11] for further details on the feature sets.

B. Temporal Modeling of Acoustic Features

Recurrent neural networks (RNNs) are a special class of deep neural networks that can learn temporal characteristics from time-series data. In the context of this work, RNNs are used to model temporal variations in speech paralinguistics within utterances of subjects from the two groups. Although there are various types of recurrent models, the two most popular structures include the Long-Short Term Memory and the Gated Recurrent Unit. Both of these provide various improvements over the legacy recurrent neural network structure (also known as *vanilla RNNs*) which has been difficult to train historically [23], [24]. This study makes use of four variations of the LSTM and one variation of the GRU for performing the temporal modeling.

1) Long Short Term Memory (LSTM): The LSTM was invented by Hochreiter and Schmidhuber [25] aiming to alleviate the vanishing/exploding gradient problem of vanilla RNNs. An LSTM cell, shown in Fig. 2, consists of four interacting layers which include *forget gate*, *input gate*, *update layer*, and *output gate*. Each of these can be considered fully connected networks in their own right, and are therefore trainable. These layers enable the LSTM cell to learn temporal patterns by managing hidden state h_t, cell state C_t, and the output of LSTM cell y_t.

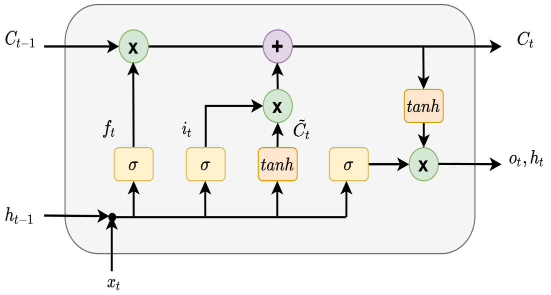

Fig. 2. Illustration of an LSTM Cell (Adopted from [26]).

The first layer in an LSTM cell is the forget gate, which is responsible for identifying parts of the previous cell state (C_{t-1}) that should be removed during the forthcoming update. If this is the first time step of the training process, the previous cell state is generated through random initialization. Next starts the process of updating the cell state with new information. Here, the input gate first identifies the location of values within the cell state which should be updated (and by how much). Then, a candidate vector of cell state values is prepared by passing the combination of input and hidden state through a *tanh* activation to squash their values between -1 and 1. Now, the cell state is updated by the summation of two products: (1) the output of forget gate and the cell state from the previous time-step ($f_t * C_{t-1}$) and (2) the input gate output and the candidate cell state ($i_t * \tilde{C}_t$). The output gate makes decisions about the parts of the cell state which should be produced as the output of the LSTM cell. Finally, the hidden state of the LSTM cell is updated for the next time-step by multiplying the cell output by the *tanh* squashed cell state. Mathematically, the process flow within the LSTM can be summarized as follows, with x_t representing the input N-dimensional acoustic features:

$$f_t = \sigma \left(W_f \cdot [h_{t-1}, x_t] + b_f \right)$$

$$i_t = \sigma \left(W_i \cdot [h_{t-1}, x_t] + b_i \right)$$

$$\tilde{C_t} = \sigma \left(W_c \cdot [h_{t-1}, x_t] + b_c \right)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C_t}$$

$$o_t = \sigma \left(W_o \cdot [h_{t-1}, x_t] + b_o \right)$$

$$h_t = o_t * tanh \left(C_t \right)$$

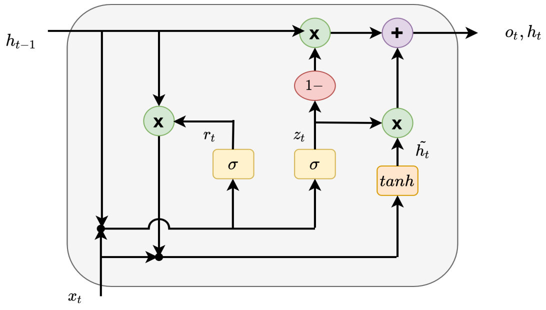

Fig. 3. Illustration of an GRU Cell (Adapted from [26]).

A bidirectional LSTM (BLSTM) is similar to an LSTM network, except that instead of the just processing the input in the forward direction for one LSTM cell, a BLSTM has another cell side by side which is fed the input in a reversed manner. By doing this, a BLSTM is able to learn from future occurrences in sequences, and is able to form more complex models than the simple LSTM. BLSTMs have been used in a variety of applications successfully (automatic speech recognition [27], voice conversion [28], etc.).

2) Gated recurrrent unit: The GRU is a temporal sequence network proposed by Cho et al. [29] with a cell structure that is more simplified than LSTM's cell structure and also has a fewer number of parameters. The GRU cell, as shown in Fig. 2, achieves the reduced complexity by combining the functions of forget and input gates from the LSTM cell into a single update gate.

Mathematically, the process flow within the GRU cell can be summarized as:

$$z_t = \sigma \left(W_z \cdot [h_{t-1}, x_t] + b_z \right)$$

$$r_t = \sigma \left(W_r \cdot [h_{t-1}, x_t] + b_r \right)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h_t}$$

3) Temporal models for recognizing dementia speech: The process of determining a particular neural network architecture is usually a trial-and-error process based on cross-validation and guided by intuition. To overcome this, we investigate the efficacy of five temporal models to identify the best performing one for the task at hand. A summary of the structure of these models is provided in Table II.

TABLE II. A SUMMARY OF FIVE SEQUENCE MODELS USED FOR LEARNING TEMPORAL CHARACTERISTICS OF SPEECH

Model ID	Model Summary
Model 1	LSTM(N) + LSTM(2*N) + LSTM(N) + Dense(N) + Dense(2)
Model 2	GRU(N) + GRU(2*N) + GRU(N) + Dense(N) + Dense(2)
Model 3	BLSTM(N) + BLSTM(N) + Dense(N) + Dense(2)
Model 4	LSTM(N) + LSTM(2*N) + LSTM(N) + Attention + Dense(N) + Dense(2)
Model 5	BLSTM(N) + BLSTM(N) + Attention + Dense(N) + Dense(2)

The first model consists of the three-layer stacked LSTM with two dense layers, including the final layer which serves as the classifier. The first and last LSTM layer has recurrent units equal to the dimensionality (N) of the input acoustic features. For example, since ComParE LLDs have a dimensionality of 65, therefore the first and last LSTM layers in `Model 1` have 65 units. The middle layer is twice the size of the first and the last LSTM layers. We decided to use these settings assuming that a larger number of recurrent units may assist in learning complex temporal patterns from acoustic features. However, this is not straightforward, since the limited amount of training data may impact the learning ability of deeper LSTM models. Moreover, a dense layer was added to investigate whether the addition of a fully connected layer can aid in learning a more meaningful representation from the acoustic embedding learned by networks. `Model 2` is identical to `Model 1` except that it consists of GRU blocks rather than LSTM cells.

`Model 3` is based on bi-directional LSTMs (BLSTM). It consists of two LSTM layers, each with the number of units equal to the dimensionality of the input acoustic LLDs and two dense layers. Here, we used smaller number of BLSTM layers, since BLSTMs naturally train two LSTMS, one in forward direction and the other - in reverse direction of the sequence.

`Model 4` and `Model 5` are similar to `Model 1` and `Model 3` except that an attention layer [30] is added to the networks. The attention layer takes into consideration the affect of long-term speech characteristics that may be learned by the LSTMs and have been found to improve the LSTM performance [31].

C. Static Modelling of Acoustic Features

Whereas temporal models seek to learn a global representation for speech recordings by explicitly considering inter-frame changes, static modelling generates a global representation of speech through feature aggregation. The simplest method of static modelling is to compute functionals of descriptive statistics for acoustic features. While this method may appear trivial, it has often produced state-of-the-art classification performance for paralinguistics tasks [32], [33]. Further, we found static modelling of acoustic features to be useful in our previous research on recognition of perceived trustworthiness [**?**]. Therefore, we benchmark the classification performance of temporal models against the static models for the task of speech-based detection of dementia.

V. Experiments and Results

The classification experiments were conducted using the stratified 5-fold cross validation method. Stratification ensures that the distribution of labels in the training partition was matched by the distribution in test partition. The performance of temporal models was benchmarked against the static models using the same LLD feature sets under the same cross-validation settings. Two types of classifiers, the support vector machine classifier (SVC) and random forest classifier (RFC), are used for classification with default settings as given in the Scikit-learn toolkit [34]. The results are presented in Table III, where it can be seen the best performing model, ComParE-SVC, yields an accuracy of 66.92%.

TABLE III. Summary of Classification Results for Static Models for IS10-Paralinguistics, ComParE, and eGeMAPS Feature Sets

Feature set	***Accuracy (%)***	
	SVC	**RF**
IS10-Paralinguistics	**61.92**	58.85
ComParE	**66.92**	61.92
eGeMAPS	57.31	**62.31**

TABLE IV. Summary of Classification Results for Temporal Models using IS10-Paralinguistics Feature Set

Model Name	**Accuracy (%)**
Model 1	70.02
Model 2	64.35
Model 3	71.79
Model 4	72.50
Model 5	**74.55**

TABLE V. Summary of Classification Results for Temporal Models using ComParE Feature Set

Model Name	**Accuracy (%)**
Model 1	64.87
Model 2	60.02
Model 3	67.14
Model 4	65.34
Model 5	**69.60**

TABLE VI. Summary of Classification Results for Temporal Models using eGeMAPS Feature Set

Model Name	**Accuracy (%)**
Model 1	66.50
Model 2	64.64
Model 3	68.93
Model 4	71.70
Model 5	**73.84**

In Table IV, we summarize the classification performance of the static and temporal models for IS10-Paralinguistics acoustic features. As can be seen, the best performance for the static modelling of features is 61.92% whereas the best performing temporal model, `Model 5` which uses BLSTM-Attention, achieves an accuracy of 74.55%, closely followed by `Model 4` which is based on LSTM-Attention achieves an accuracy of 72.50%. While these results are preliminary, they indicate that attention mechanism can assist in improving the classification models.

The classification performance using ComParE feature set has been summarized in Table V. Here, the static modelling achieves the best performance with a classification accuracy of 62.31%. In contrast, best model amongst the temporal modelling approaches achieves an accuracy of 69.60%. Interestingly, the performance of the best sequence modelling result is significantly lower when compared to the result of the previous experiment. We suggest this is because the dimensionality of ComParE is much larger than that of IS10-Paralinguistic features (130 versus 76), and there are not enough examples in the dataset to adequately train temporal models with ComParE features.

Finally, Table VI shows the summary of classification performance for the static and temporal modelling using eGeMAPS features. As can be seen, the best performing model for the static features is `eGeMAPS-RF` which achieves an accuracy of 62.31%, whereas the best performing model amongst temporal models is `Model 5` that provides an accuracy of 73.84%. The second placed temporal model, `Model 4`, achieves an accuracy of 71.70%. Both these results are more superior than the accuracy achieved through the static-modelling.

VI. Discussion

The results presented above lead to some interesting observations. Firstly, we note that the temporal models provide higher classification performance than the static models. A caveat to this is the performance of the temporal models with ComParE features which offer relatively smaller improvements than IS10-Paralinguistics and eGeMAPS feature set. We suggest this is due to the large dimensionality of ComParE LLDs – 130 for ComParE versus 76 for IS10-Paralingusitics and 23 for eGeMAPS LLDs, respecively. Another observation is that attention mechanism contributed to the consistently improved classification by the temporal models. We believe this is because attention mechanism assists the temporal models in focusing on time-dependent charactersitics of acoustic LLDs that are unique for individuals with dementia.

Table VII summarizes the classification results of top-3 best performing models, where one can note that all of these models are based on the temporal modelling with attention mechanism, and two out of these make use of the IS10-Paralinguistic feature set. As can be seen, a significant improvement in classification accuracy is achieved when compared to the best performing statistic model (which achieved 66.92%).

TABLE VII. Summary of Top-3 Temporal Models

Model	**Acoustic Feature**	***Accuracy (%)***
Model 5	IS10-Paralinguistics	74.55
Model 5	eGeMAPS	73.84
Model 4	IS10-Paralinguistics	72.50

VII. Conclusion

In this paper, we investigated the efficiency of the temporal modelling vs. static modeling of speech acoustics for detection of individuals with dementia. We benchmarked the proposed temporal models with the static models of acoustic features. Experimental results showed that the temporal modelling is a more effective approach for the intended classification task, revealing the best-case accuracy of 74.55% in a 5-fold cross-validation setup. This accuracy may not be sufficient to support a medical diagnosis; however it is sufficiently high to conduct a low-cost rapid screening for dementia that could be followed up by a professional assessment. The proposed acoustic speech classification could be used either alone or in combination with transcript analysis, questionnaires, and other standard screening techniques.

Our future work will investigate application of other deep learning models and integration of dynamic and static approaches into a combined decision-making system.

References

[1] World Health Organization, "The Epidemiology and Impact of Dementia - Current State and Future Trends," Tech. Rep., 2018. [Online]. Available: https://www.who.int/health-topics/dementia

[2] Stanford Healthcare, "Causes of Dementia," 2020. [Online]. Available: https://stanfordhealthcare.org/medical-conditions/brain-and-nerves/dementia/causes.html

[3] Alzheimers Association, "What Is Dementia?" 2020. [Online]. Available: https://www.alz.org/alzheimers-dementia/what-is-dementia

[4] K. Ritchie, I. Carriere, L. Su, J. T. O'Brien, S. Lovestone, K. Wells, and C. W. Ritchie, "The midlife cognitive profiles of adults at high risk of late-onset Alzheimer's disease: The PREVENT study," *Alzheimer's and Dementia*, vol. 13, no. 10, pp. 1089–1097, 2017.

[5] M. Mortamais, J. A. Ash, J. Harrison, J. Kaye, J. Kramer, C. Randolph, C. Pose, B. Albala, M. Ropacki, C. W. Ritchie, and K. Ritchie, "Detecting cognitive changes in preclinical Alzheimer's disease: A review of its feasibility," *Alzheimer's and Dementia*, vol. 13, no. 4, pp. 468–492, 2017.

[6] C. Laske, H. R. Sohrabi, S. M. Frost, K. Lopez-De-Ipina, P. Garrard, M. Buscema, J. Dauwels, S. R. Soekadar, S. Mueller, C. Linnemann, S. A. Bridenbaugh, Y. Kanagasingam, R. N. Martins, and S. E. O'bryant, "Innovative diagnostic tools for early detection of Alzheimer's disease," *Alzheimer's and Dementia*, vol. 11, no. 5, pp. 561–578, 2015.

[7] V. Boschi, E. Catricala, M. Consonni, C. Chesi, A. Moro, and S. F. Cappa, "Connected speech in neurodegenerative language disorders: A review," *Frontiers in Psychology*, vol. 8, pp. 1–21, 2017.

[8] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's Dementia Recognition through Spontaneous Speech: The ADReSS Challenge," in *INTERSPEECH (to appear)*, 2020, pp. 1–5.

[9] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," in *ACM International Conference on Multimedia*, 2013, pp. 835–838.

[10] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native Language," in *INTERSPEECH*, 2016, pp. 2001–2005.

[11] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "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, 2016.

[12] 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.

[13] 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.

[14] 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, 2015.

[15] D. Yu and L. Deng, *Automatic speech recognition.* Springer, 2016.

[16] B. Mirheidari, D. Blackburn, T. Walker, M. Reuber, and H. Christensen, "Dementia detection using automatic analysis of conversations," *Computer Speech and Language*, vol. 53, pp. 65–79, 2019.

[17] C. Elsey, P. Drew, D. Jones, D. Blackburn, S. Wakefield, K. Harkness, A. Venneri, and M. Reuber, "Towards diagnostic conversational profiles of patients presenting with dementia or functional memory disorders to memory clinics," *Patient Education and Counseling*, vol. 98, no. 9, pp. 1071–1077, 2015.

[18] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Muller, and S. Narayanan, "Paralinguistics in speech and language — State-of-the-art and the challenge," *Computer Speech and Language*, vol. 27, no. 1, pp. 4–39, 2013.

[19] Z. S. Syed, K. Sidorov, and D. Marshall, "Depression Severity Prediction Based on Biomarkers of Psychomotor Retardation," in *ACM*

International Workshop on Audio/Visual Emotion Challenge (AVEC), 2017, pp. 37–43.

[20] ——, "Automated Screening for Bipolar Disorder from Audio/Visual Modalities," in *ACM International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2018, pp. 39–45.

[21] S. Amiriparian, A. Baird, S. Julka, A. Alcorn, S. Ottl, S. Petrovic, E. Ainger, N. Cummins, and B. Schuller, "Recognition of Echolalic Autistic Child Vocalisations Utilising Convolutional Recurrent Neural Networks," in *INTERSPEECH*, 2018, pp. 2334–2338.

[22] S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, M. Christian, S. Language, P. Group, D. Telekom, and A. G. Laboratories, "The INTERSPEECH 2010 Paralinguistic Challenge," in *INTERSPEECH*, 2010, pp. 2794–2797.

[23] Y. Bengio, P. Frasconi, and P. Simard, "Problem of learning long-term dependencies in recurrent networks," in *IEEE International Conference on Neural Networks*, 1993, pp. 157–166.

[24] Y. Bengio, P. Simard, and P. Frasconi, "Learning Long-Term Dependencies with Gradient Descent is Difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[25] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[26] C. Olah, "Understanding LSTM Networks," 2015. [Online]. Available: https://colah.github.io/posts/2015-08-Understanding-LSTMs

[27] A. Zeyer, P. Doetsch, P. Voigtlaender, R. Schlüter, and H. Ney, "A comprehensive study of deep bidirectional lstm rnns for acoustic modeling in speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2462–2466.

[28] H. Ming, D. Huang, L. Xie, J. Wu, M. Dong, and H. Li, "Deep bidirectional lstm modeling of timbre and prosody for emotional voice conversion," 2016.

[29] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," *arXiv:1406.1078*, pp. 1–15, 2014.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017.

[31] Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou, and B. Schuller, "Speech emotion classification using attention-based lstm," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1675–1685, 2019.

[32] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom, A. S. Warlaumont, G. Hidalgo, S. Schnieder, C. Heiser, W. Hohenhorst, M. Herzog, M. Schmitt, K. Qian, Y. Zhang, G. Trigeorgis, P. Tzirakis, and S. Zafeiriou, "The INTERSPEECH 2017 Computational Paralinguistics Challenge: Addressee, Cold and Snoring," in *INTERSPEECH*, 2017, pp. 1–5.

[33] B. W. Schuller, S. Steidl, A. Batliner, P. B. Marschik, H. Baumeister, F. Dong, S. Hantke, F. Pokorny, E.-M. Rathner, K. D. Bartl-Pokorny, C. Einspieler, D. Zhang, A. Baird, S. Amiriparian, K. Qian, Z. Ren, M. Schmitt, P. Tzirakis, and S. Zafeiriou, "The INTERSPEECH 2018 Computational Paralinguistics Challenge: Atypical and Self-Assessed Affect, Crying and Heart Beats," in *INTERSPEECH*, 2018, pp. 1–5.

[34] 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.