A Comparison of Acoustic and Linguistics Methodologies for Alzheimer's Dementia Recognition

Conference Paper · October 2020

DOI: 10.21437/Interspeech.2020-2635

11 authors, including:

Nicholas Cummins
Universität Augsburg
174 PUBLICATIONS **4,118** CITATIONS
SEE PROFILE

Yilin Pan
The University of Sheffield
13 PUBLICATIONS **105** CITATIONS
SEE PROFILE

Zhao Ren
Forschungszentrum L3S
59 PUBLICATIONS **1,015** CITATIONS
SEE PROFILE

Julian Fritsch
Idiap Research Institute
7 PUBLICATIONS **96** CITATIONS
SEE PROFILE

A Comparison of Acoustic and Linguistics Methodologies for Alzheimer's Dementia Recognition

Nicholas Cummins[1], *Yilin Pan*[2], *Zhao Ren*[1], *Julian Fritsch*[3,4], *Venkata Srikanth Nallanthighal*[5], *Heidi Christensen*[2], *Daniel Blackburn*[6], *Björn W. Schuller*[1,7], *Mathew Magimai.-Doss*[3], *Helmer Strik*[8], *Aki Härmä*[5]

[1]Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[2]Department of Computer Science, University of Sheffield, UK
[3]Idiap Research Institute, Martigny, Switzerland
[4]École polytechnique fédérale de Lausanne (EPFL), Switzerland
[5]Philips Research, Eindhoven,The Netherlands
[6]Sheffield Institute for Translational Neuroscience (SITraN), University of Sheffield, UK
[7]GLAM – Group on Language, Audio, & Music, Imperial College London, UK
[8]Centre for Language Studies (CLS), Radboud University Nijmegen

nicholas.cummins@ieee.org

Abstract

In the light of the current COVID-19 pandemic, the need for remote digital health assessment tools is greater than ever. This statement is especially pertinent for elderly and vulnerable populations. In this regard, the INTERSPEECH 2020 Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) Challenge offers competitors the opportunity to develop speech and language-based systems for the task of Alzheimer's Dementia (AD) recognition. The challenge data consists of speech recordings and their transcripts, the work presented herein is an assessment of different contemporary approaches on these modalities. Specifically, we compared a hierarchical neural network with an attention mechanism trained on linguistic features with three acoustic-based systems: (i) Bag-of-Audio-Words (BoAW) quantising different low-level descriptors, (ii) a Siamese Network trained on log-Mel spectrograms, and (iii) a Convolutional Neural Network (CNN) end-to-end system trained on raw waveforms. Key results indicate the strength of the linguistic approach over the acoustics systems. Our strongest test-set result was achieved using a late fusion combination of BoAW, End-to-End CNN, and hierarchical-attention networks, which outperformed the challenge baseline in both the classification and regression tasks.

Index Terms: Alzheimer's Disease, Bag-of-Audio-Words, Convolutional Neural Network, Siamese Network, Hierarchical Neural Network, Attention Mechanisms

1. Introduction

According to the *World Health Organisation* (WHO), dementia is a major cause of disability in the elderly population worldwide, with at least 10 million new cases reported every year [1]. Alzheimer's Disease (AD) is the most common cause of dementia [1, 2] and is a major public health concern, with considerable associated socio-economic costs [2]. Therefore, there is an urgent need for early diagnosis systems in order to promote timely and optimal management. The current coronavirus disease 2019 (COVID-19) pandemic accelerates this need; people living with dementia are at an increased risk of infection due to an inability to comprehend, recall and follow hygiene and procedures [3].

Declines in speech and language are regarded as key early markers of AD [4]. However, sparse and heterogeneous data sets are limiting the impact of research in this area. The *Alzheimer's Dementia Recognition through Spontaneous Speech* (ADReSS) challenge aims to address this issue by supplying a new AD speech corpus on which competitors perform two different recognition tasks [5]. The database consists of 54 participants with AD and 54 matched controls. The first task is the 2-class classification between the AD and non-AD samples. The second task is a regression task predicting the score of the *Mini-Mental State Examination* (MMSE) [6] of a speaker.

Herein, we present the ***T**raining Network on **A**utomatic **P**rocessing of **PA**thological **S**peech* (TAPAS) – a Horizon 2020 Marie Skłodowska-Curie Actions Innovative Training Network European Training Network – approach to these two tasks. As both acoustic- and linguistics-based systems have shown promise in the identification of AD, the latter particularly so, we explore the efficacy of combining information gained from these different combinations of state-of-the-art approaches.

Based on previous works that demonstrated their suitability in related tasks, we utilise three different acoustic-based systems. The first, a *Bag-of-Audio Words* (BoAW) system [7] has been successfully applied for other speech-health recognition tasks, e. g., [8, 9]. Based on results achieved in [10], we also test a Siamese network [11]. Finally, building on [12], we include an End-to-End, raw waveform, *Convolutional Neural Network* (CNN). To the best of the authors' knowledge, this is the first time these three systems have been used in AD recognition.

We compare and combine these acoustic systems, with a linguistic system that utilises *Global Vectors* (GLoVe) word embeddings [13] and a hierarchical attention neural network [14]. The strength of combining word-embeddings and hierarchical attention networks has been demonstrated across a range of natural language processing (NLP) tasks [15], including AD detection [16, 17]. Given that linguistic features have, in general, shown stronger performances in AD detection tasks [4, 18], we regard this system as our gold standard, and investigate if, (i) a combination of our acoustic systems can match performance with the linguistic systems and, (ii), if the acoustic systems provide complementary information to the linguistic system.

2. Methodology

This section introduces the acoustic and linguistic systems considered in our contribution to the ADReSS challenge.

2.1. Bag-of-Audio-Words

Bag-of-Audio-Words (BoAW) [7] features have been applied for a range of speech-based recognition tasks, including cold and flu detection [8] and level of pain evaluation [9]. BoAW involves the quantisation of acoustic *low-level descriptors* (LLDs), where each frame-level LLD vector is assigned to an audio word from a previously learnt codebook. Typically, the codebook is formed form LLDs extracted from the training partition of a dataset. The subsequent quantisation, undertaken by counting the number of assignments for each audio word, generates a sparse histogram representation of a given speech file. The openXBOW [7] is an open-sourced toolbox for the formation of BoAW features, it has been widely utilised such as in INTERSPEECH Computational Paralinguistics Challenges (COMPARE) [19].

In the formation of BoAW features, LLD vectors are first extracted from the speech files. In this work, three LLDs feature representations are generated using openSMILE [20]: Mel-Frequency Cepstral Coefficient (MFCC), log-Mel, and the COMPARE acoustic feature set [21]. These three feature representations have previously been shown to be suitable for AD recogniton [4, 5, 22] and their bagged representations have performed well in previous studies, especially in health-based tasks [23]. Therefore, the three BoAW representations have promise as effective representations of AD recognition.

2.2. Siamese Network

Inspired by the success of Siamese networks in related tasks [10, 24–26], we investigate this paradigm for the task of AD recognition. A core advantage of Siamese networks is the associated *contrastive loss function* that encourages intra-class compactness and inter-class separability [27]. During training, information from segments of recordings belonging to the same condition (AD speech or healthy speech) is pulled together using contrastive loss, while information relating to segments of recordings from different conditions are pushed away from each other.

Formally we define the contrastive loss L_c as:

$$L_c(\boldsymbol{x}_1, \boldsymbol{x}_2, y) = (1-y) \cdot D(\boldsymbol{x}_1, \boldsymbol{x}_2)^2 + y \cdot \max(0, \alpha_c - D(\boldsymbol{x}_1, \boldsymbol{x}_2))^2, \quad (1)$$

where $y = 1$ if the embeddings $\boldsymbol{x}_1$, and $\boldsymbol{x}_2$ are from different conditions and should thus be distant, and $y = 0$ when $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are from the same condition and thus should be close. Additionally, D denotes the Euclidean distance and α_c is the margin which we want to obtain between the two different conditions.

2.3. End-to-End Convolutional Neural Network

We also propose modelling AD's speech in an end-to-end manner, utilising raw waveform based CNNs. This framework was been successfully applied to tasks such as emotion recognition [28], speaker verification [29], gender identification [30], or depression detection [12]. Exploiting this paradigm, we can capture information related to different speech production mechanisms by modifying the initial kernel width (kW) parameter [29, 31]. Setting $kW = 300$ covers a signal length of approximately 20 ms (segmental) allows the first convolution layer to model voice-source-related information. Alternatively, by setting $kW = 30$ covers a signal of approximately 2 ms of length (sub-segmental), encouraging the first convolution layer into tending to capture vocal tract information, such as formants.

In order to verify the importance of changes in fundamental frequency, we also investigated using zero-frequency filtered (ZFF) signals [32]. Taking inspiration from a recent paper showing that voice source related information related to depression can be modelled with CNNs [12], the filtered signals are fed to the same network applied to classification and regression tasks.

2.4. Hierarchical Attention Network

We implement a *bi-directional Hierarchical Attention Network* (bi-HANN) as our linguistic system. This choice was motivated by the success of bi-HANNs in other AD recognition tasks [16, 17]. This approach is a two-stage system which operates at the word- and sentence-level [14]. In our model, w_{it} with $t \in [1, T]$ and $i \in [1, L]$ is used to represent the tth word in ith sentence. Each word w_{ij} is encoded into a fixed dimensional vector x_{ij} by a pre-trained embedding matrix W_e. The choice word embedding matrix is a trainable parameter in the model.

To extract word-level characteristic patterns from the variable-length sequence, a *bidirectional long short-term memory* (bi-LSTM) is applied on the word vectors, followed by an attention mechanism. After obtaining the sentence representation s_i, a further bi-LSTM layer extracts sentence-level information extraction. Given a sentence vector s_i, this action generates a transcript representation v with a similar structure as for the word level model. Finally, a dense layer with a *sigmoid* function is applied for classification on the transcript representation. See [17] for further information on this paradigm.

3. Experimental Setup

This section introduces the ADReSS AD dataset, as well as the key outlines the key experimental settings associated with our four AD recognition systems.

3.1. Database

The speech data and transcripts used in this work were provided by the ADReSS challenge organisers [5]. The speech data contains both full speech files and segmented speech chunks. The segmented chunks, used to set the challenge baseline [5] were generated by the organisers applying a log-energy threshold-based voice activity detector. The BoAW and End-to-End systems utilised these chunks, while the Siamese network exploits the full recordings. The transcripts contain the linguistic content of an interviewer and a participant, as well as other related annotations. We, therefore, pre-processed all the raw transcripts to remove all content unrelated to the spoken content of the participant and used the remaining information as input to the bi-HANN. For the sake of brevity, the demographics and characteristics of the data set are not given here. The interested reader is referred to [5] for these details.

3.2. Bag-of-Audio-Words

The extraction of the three LLDs representations mentioned in 2.1 is described below. Both MFCC and log-Mel LLDs are extracted with a frame size of 0.025 s and a step size of 0.01 s. The MFCC LLDs consist of MFCC 1-14 and the corresponding delta regression coefficients, leading to 28-dimensional MFCC LLDs. The log-Mel LLD feature set contains 64-band log-Mel frequencies and corresponding 64 delta regression coefficients. The 130 dimensional COMPARE LLDs [21] were obtained by the OPENSMILE configure file *ComParE_2016.conf*.

Next, the three LLD feature sets were fed into the BoAW to extract representations for each audio sample. The input LLDs are split into two subsets, in order to train a separate codebook in each subset. We then quantise 14 LLDs for MFCC, 64 for log-Mel, and 65 for ComParE features for both subsets. The number of word-assignments was set as 10 for all three feature sets. Then, the optimal codebook size was searched in $\{65, 125, 250, 500\}$. Finally, the extracted BoAW features were then fed into a linear Support Vector Machine (SVM) for classification or regression. The complexity hyperparameter in the SVM is optimised from the setting of $\{1e^{-6}, 1e^{-5}, 1e^{-4}, 1e^{-3}, 1e^{-2}, 1e^{-1}, 1\}$.

3.3. Siamese Networks

This model generates embeddings using a deep Siamese neural network consisting of convolutional layers. The network was trained using a contrastive loss between the two different conditions (Section 2.2). Note, as the Siamese network and contrastive loss function are not suited to regression analysis, we only present use of this system in the classification task.

As an input, the model used either 8-second or 16-second segments extracted from the full, rather than the chunked audio recordings (Section 3.1). The segments were generated using a sliding window of 2 seconds over the recordings and were represented as log-Mel spectrograms. Our deep Siamese network consists of two CNNs to extract embeddings one from each of the two inputs. The encoded embeddings are then concatenated and fed into a fully connected network to estimate their similarity. Specifically, each CNN has 4 convolutional layers, each of which is followed by rectified linear unit (ReLU) activations, and batch normalisation. After the embeddings from the two CNNs are extracted, they are concatenated and fed into a 2-layer Fully Connected Network with each layer followed by ReLU activation. The final layer uses a sigmoid activation function to squash the output value between 0 and 1, which is regarded as the similarity value.

3.4. End-to-End Convolutional Neural Network

Raw waveform CNN networks typically consist of an initial filter stage followed by a classification stage. Our proposed network has four convolution layers, kernel widths *30-10-4-3* for subsegmental modelling, and *300-5-4-3* for segmental modelling (Section 2.3). Convolution layers are followed by maximum pooling and ReLU activations. The final stage of the network is a multilayer perceptron. At the output, the classification network predicts a probability for AD using a sigmoid function, while the output is a linear value for the regression model. In both cases, this is a per frame action. These frame-level values are then averaged to get per-utterance posterior probabilities.

The input to the CNNs, w_{seq}, is a 250 ms length speech segment, shifted by 10 ms. We randomly-initialised CNNs with a batch-size of 256 and employ a cross-entropy cost function or mean squared error for the two tasks, respectively. We opted for a decaying learning schedule which halves the learning rate between 10^{-3} and 10^{-7} whenever the validation loss stops reducing. In initial testing, we observed that a combination of *ZFF* with subsegmental modelling was better suited to the classification task. In contrast, the combination of *ZFF* with segmental modelling was better suited to the regression task. Herein, *ZFF* denotes this combination.

3.5. Hierarchical Attention Network

Only the transcripts that corresponded to participants are used for the bi-HANN model (Section 3.1). GloVe 100-dimensional word vectors trained on Wikipedia 2014 and Gigaword-5 data were taken as our pre-trained embedding matrix [13]. The bi-HANN was trained on a fixed number of epochs (20) and evaluated on the development set at each epoch. Batch size was set to 20 and the best model selected via the F_1-score on the training set. The number of LSTM units was set to 100, and the dense layer dimension in word-level was set as 50. For the attention layer's dimension, both the sentence and word level is set to 30. The sentence length was set to 30, and we zero-padded the shorter sentences. The sentence numbers in a transcript were set to 30, with zero-padding used on the shorter transcript. We opted for *Adam* optimisation with a learning rate of $1e^{-5}$. Dropout was applied after all the functional layers with 0.3 dropout rate.

We compare the *bi-HANN* with two simplified linguistics systems, a *bi-LSTM* and *bi-LSTM with attention* (bi-LSTM-Att). These models follow the same parameters setting as in the *biHANN*. The maximum word number for each transcript is 200, with zero-padding being applied if the word number is less than this amount. Dropout layers are adopted after the LSTM layer and attention, and dense layers.

Table 1: *A comparison of the proposed approaches on the ADReSS Challenge training set. Results are the average performance across a nine-fold cross-validation step up.*

Approach			**Acc.**	**F1**	**RMSE**
BoAW	MFCC	65	.611	.604	7.03
		125	**.630**	**.623**	7.05
		250	.602	.593	7.00
		500	.620	.610	7.17
	LogMel	65	.565	.540	7.18
		125	.556	.526	**6.97**
		250	.537	.522	7.15
		500	.556	.544	7.03
	ComParE	65	.620	.601	7.04
		125	.593	.582	7.04
		250	.574	.556	7.17
		500	.574	.567	7.13
	Fusion	–	**.639**	.639	6.99
SiameseNet	LogMel	8 s	.586	.693	–
		16 s	**.628**	**.731**	–
End-to-End	Raw seg		.713	.762	8.89
	ZFF		**.741**	**.780**	7.58
Linguistic	bi-LSTM		.694	.736	5.99
	bi-LSTM-Att		**.842**	**.842**	5.49
	bi-HANN		.827	.826	**4.86**
Fusion	Maj. / Wt.		.850	.855	4.91
Fusion	bi-LSTM-Att		.887	.887	7.73
Fusion	bi-HANN		.831	.829	7.64

3.6. Evaluation Metrics

As per the challenge organisers [5], we report our results in terms of accuracy and F_1-score for the classification score, and *root mean squared error* (RMSE) for the MMSE prediction task. We divided the 108 speaekers in the training set into 9 folds of 12 speakers[1] and report the average of each score across each fold in the results section.

[1]Partitioning of folds available on request

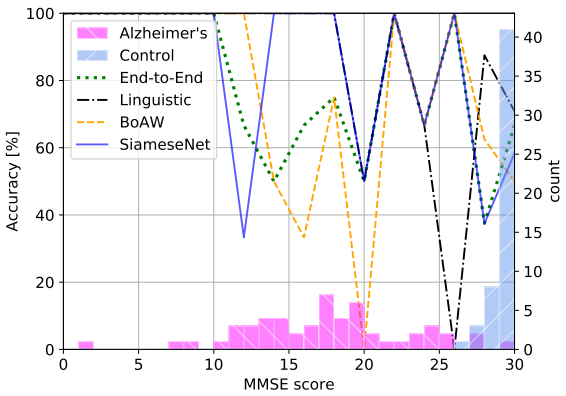

Figure 1: *Accuracy per MMSE score of our for best systems on the development set, together with a histogram of MMSE scores.*

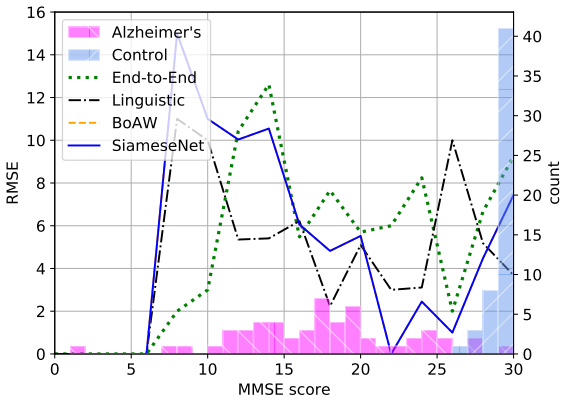

Figure 2: *RMSE per MMSE score of our for best systems on the development set, together with a histogram of MMSE scores.*

4. Results and Discussion

4.1. Training Set Results

As expected, the linguistic systems outperforms the acoustic systems (Table 1). The *bi-HANN* system achieves the strongest result on the regression task; however, the simpler *bi-LSTM-Att* system achieves the strongest performances on the classification task. This result does not match with similar systems in the literature [17]. We speculate the more even performances between the *bi-HANN* and *bi-LSTM-Att* systems are due to the smaller size of the ADReSS database. The end-to-end CNNs produce the strongest performance of the acoustic systems on the classification task, highlighting the benefits of self-learning features. The inclusion of the *ZFF* signals improves the performance of this set-up, indicating the importance of fundamental frequency in AD recognition tasks. Finally, the *BoAW-logMel-C125* gains the lowest RMSE of our acoustic systems; verifying the strength of this feature representation in paralinguistic tasks [23].

Figure 1 and Figure 2 show accuracy and RMSE per MMSE, respectively, for the the best performing systems from each group on the training set. In terms of accuracy, none of the systems in Figure 1 show any consistency. Whereas, in terms of RMSE, we observe high errors at low MMSE values and another peak around 26, where control and AD histograms start overlapping.

The late fusion between the best-performing systems from each grouping did not improve system performance beyond the linguistic only approaches (Table 1). This approach adopted a majority voting for the classification task or a weighted average for the regression task. However, in the classification task, when fusing the *bi-LSTM-Att* and *ZFF* systems, we were able to improve on the performance of the bi-LSTM-Att system. This approach exploited the learnt representations from the second to last layer of the *ZFF* CNN. These features were concatenated with the attention output of the bi-LSTM attention layer and the network trained as per (Section 2.4). However, this feature fusion approach was not as well suited to the regression task.

Table 2: *A comparison of the best performing approaches from Table 1 on the ADReSS Challenge training set*

Approach	**Acc.**	**F1**	**RMSE**
Baseline [5]	.625	.620	6.14
BoAW	.563	.561	6.88
BoAW fusion (3-best)	.625	.625	6.45
SiameseNet	**.708**	**.708**	–
End-to-End	.667	.664	6.75
bi-LSTM-Att	.813	.812	4.66
bi-HANN	.729	.726	4.74
Fusion Feat. (bi-LSTM-Att)	.771	.766	5.62
Fusion Feat. (bi-HANN)	.813	.810	6.65
Fusion Maj./ W-avg (3-best)	**.852**	**.854**	**4.65**

4.2. Test Set Results

The SiameseNet performs the strongest out of the acoustic systems in the classification task (Table 2). Interestingly, despite their stronger performance in the classification task, none of the acoustic systems trailed on the test set out-performs the regression baseline. The *bi-LSTM-Att* system was our strongest standalone system, highlighting the strength of considering linguistics in AD recognition tasks. The benefits of fusion are more apparent in the test set, with our best result being achieved through a majority vote (classification) / weighted average (fusion) of the *BoAW-MFCC-C125* (classification) / *BoAW-logMel-C125* (regression), *ZFF*, and *bi-LSTM-Att* systems. This set-up achieves an accuracy of .852 and an RMSE of 4.65.

5. Conclusions

This paper described the TAPAS Training Network approach to the INTERSPEECH 2020 ADReSS challenge. We compared and combined information from four different speech-based Alzheimer's recognition approaches; three acoustic and one linguistic. The linguistic systems outperformed our acoustics approaches; such a result is unsurprising given a human observer generated the transcripts. Thus, they contain considerably fewer sources of noise than the audio recordings. Small gains were found when fusing acoustics and linguistics approaches. In future work, we will explore the effect of performing similar analysis when combining acoustic information with linguistics systems based on transcripts generated from an automatic speech recognition system.

6. Acknowledgements

This work was supported by the Horizon H2020 Marie Skłodowska-Curie Actions Initial Training Network European Training Network (MSCA-ITN-ETN) project under grant agreement No. 766287 (TAPAS). The four early stage researchers from the project: Yilin Pan, Zhao Ren, Julian Fritsch, and Srikanth Nallanthighal, all contributed equally to this manuscript.

7. References

[1] World Health Organization, "Towards a Dementia Plan: A WHO Guide," WHO, 2018, 82 pages. [Online]. Available: https://www.who.int/mental_health/neurology/dementia/guidelines_risk_reduction/en/

[2] Alzheimer's Association, "2017 Alzheimer's disease facts and figures," *Alzheimer's Dementia*, vol. 13, no. 4, pp. 325–373, 2017.

[3] H. Wang, T. Li, P. Barbarino, S. Gauthier, H. Brodaty, J. L. Molinuevo, H. Xie, Y. Sun, E. Yu, Y. Tang, and X. Yu, "Dementia care during COVID-19," *The Lancet*, vol. 395, no. 10231, pp. 1190–1191, 2020.

[4] 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.

[5] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's Dementia Recognition through Spontaneous Speech: The ADReSS Challenge," in *Proc. INTERSPEECH 2020*, Shanghai, China, 2020, 5 pages. [Online]. Available: https://arxiv.org/abs/2004.06833

[6] T. N. Tombaugh and N. J. McIntyre, "The Mini-Mental State Examination: A Comprehensive Review," *Journal of the American Geriatrics Society*, vol. 40, no. 9, pp. 922–935, 1992.

[7] M. Schmitt and B. Schuller, "openXBOW — Introducing the Passau Open-Source Crossmodal Bag-of-Words Toolkit," *Journal of Machine Learning Research*, vol. 18, pp. 1–5, 2017.

[8] N. Cummins, M. Schmitt, S. Amiriparian, J. Krajewski, and B. Schuller, "You sound ill, take the day off: Classification of speech affected by Upper Respiratory Tract Infection," in *Proc. EMBC 2017*. Jeju Island, South Korea: IEEE, 2017, pp. 3806–809.

[9] Z. Ren, N. Cummins, J. Han, S. Schnieder, J. Krajewski, and B. Schuller, "Evaluation of the pain level from speech: Introducing a novel pain database and benchmarks," in *Proc. ITG 2018*. Oldenburg, Germany: IEEE/VDE, 2018, pp. 56–60.

[10] S. Boelders, V. S. Nallanthighal, V. Menkovski, and A. Härmä, "Detection of Mild Dyspnea from Pairs of Speech Recordings," in *Proc. ICASSP 2020*. Barcelona, Spain: IEEE, 2020, pp. 4102–4106.

[11] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature Verification using a "Siamese" Time Delay Neural Network," in *Advances in Neural Information Processing Systems 6*, J. D. Cowan, G. Tesauro, and J. Alspector, Eds. Morgan-Kaufmann, 1994, pp. 737–744.

[12] S. P. Dubagunta, B. Vlasenko, and M. Magimai.-Doss, "Learning Voice Source Related Information for Depression Detection," in *Proc. ICASSP 2019*. Brighton, United Kingdom: IEEE, 2019, pp. 6525–6529.

[13] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. EMNLP 2014*. Doha, Qatar: ACL, 2014, pp. 1532–1543.

[14] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical Attention Networks for Document Classification," in *Proc. NAACL-HLT 2016*, San Diego, California, USA, 2016, pp. 1480–1489.

[15] A. Mallol-Ragolta, Z. Zhao, L. Stappen, N. Cummins, and B. Schuller, "A Hierarchical Attention Network-Based Approach for Depression Detection from Transcribed Clinical Interviews," in *Proc. INTERSPEECH 2019*. Graz, Austria: ISCA, 2019, pp. 221–225.

[16] J. Chen, J. Zhu, and J. Ye, "An Attention-Based Hybrid Network for Automatic Detection of Alzheimer's Disease from Narrative Speech," in *Proc. INTERSPEECH 2019*. Graz, Austria: ISCA, 2019, pp. 4085–4089.

[17] Y. Pan, B. Mirheidari, M. Reuber, A. Venneri, D. Blackburn, and H. Christensen, "Automatic Hierarchical Attention Neural Network for Detecting AD," in *Proc. INTERSPEECH 2019*. Graz, Austria: ISCA, 2019, pp. 4105–4109.

[18] B. Mirheidari, D. Blackburn, K. Harkness, T. Walker, A. Venneri, M. Reuber, and H. Christensen, "An Avatar-Based System for Identifying Individuals Likely to Develop Dementia," in *Proc. INTERSPEECH 2017*. Stockholm, Sweden: ISCA, 2017, pp. 3147–3151.

[19] B. Schuller, S. Steidl, A. Batliner, P. Marschik, H. Baumeister, F. Dong, S. Hantke, F. Pokorny, E.-M. Rathner, K. 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 & Self-Assessed Affect, Crying & Heart Beats," in *Proc. INTERSPEECH 2018*. Hyderbad, India: ISCA, 2018, pp. 122–126.

[20] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE: The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proc. ACM MM '10*. Firenze, Italy: ACM, 2010, pp. 1459–1462.

[21] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor," in *Proc. of ACM MM '13*. Barcelona, Spain: ACM, 2013, pp. 835–838.

[22] 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, 2020.

[23] N. Cummins, A. Baird, and B. Schuller, "The increasing impact of deep learning on speech analysis for health: Challenges and Opportunities," *Methods*, vol. 151, pp. 41–54, 2018.

[24] Z. Lian, Y. Li, J. Tao, and J. Huang, "Speech Emotion Recognition via Contrastive Loss under Siamese Networks," in *Proc. ASMMC-MMAC '18*. Seoul, Republic of Korea: ACM, 2018, pp. 21–26.

[25] J. Wang, Y. Qin, Z. Peng, and T. Lee, "Child Speech Disorder Detection with Siamese Recurrent Network Using Speech Attribute Features," in *Proc. INTERSPEECH 2019*. Graz, Austria: ISCA, 2019, pp. 3885–3889.

[26] M. Harandi, S. R. Kumar, and R. Nock, "Siamese Networks: A Thing or Two to Know," Data61, CSIRO, 2017. [Online]. Available: https://pdfs.semanticscholar.org/c03c/e09b419b6a15c1228e344a900d8c54bddc78.pdf

[27] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a Similarity Metric Discriminatively, with Application to Face Verification," in *ProcĊVPR '05*. San Diego, CA, USA: IEEE, 2005, pp. 539–546.

[28] G. Trigeorgis, F. Ringeval, R. Brückner, E. Marchi, M. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu Features? End-to-End Speech Emotion Recognition using a Deep Convolutional Recurrent Network," in *Proc. ICASSP 2016*. Shanghai, China: IEEE, 2016, pp. 5200–5204.

[29] H. Muckenhirn, M. Magimai-Doss, and S. Marcell, "Towards directly modeling raw speech signal for speaker verification using cnns," in *Proc. ICASSP 2018*. Calgary, AB, Canada: IEEE, 2018, pp. 4884–4888.

[30] S. H. Kabil, H. Muckenhirn, and M. Magimai-Doss, "On Learning to Identify Genders from Raw Speech Signal Using CNNs," in *Proc. INTERSPEECH 2018*. Hyderbad, India: ISCA, 2018, pp. 287–291.

[31] D. Palaz, M. Magimai.-Doss, and R. Collobert, "End-to-End Acoustic Modeling using Convolutional Neural Networks for HMM-based Automatic Speech Recognition," *Speech Communication*, vol. 108, pp. 15–32, Apr. 2019.

[32] K. S. R. Murty and B. Yegnanarayana, "Epoch Extraction From Speech Signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602–1613, 2008.