
Proceedings Paper:
Pan, Y., Mirheidari, B., Reuber, M. et al. (2020) Improving detection of Alzheimer's Disease using automatic speech recognition to identify high-quality segments for more robust feature extraction. In: Proceedings of Interspeech 2020. Interspeech 2020, 25-29 Oct 2020, Shanghai, China. International Speech Communication Association (ISCA), pp. 4961-4965. ISSN: 1990-9772.

https://doi.org/10.21437/interspeech.2020-2698

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

Improving detection of Alzheimer's Disease using automatic speech recognition to identify high-quality segments for more robust feature extraction

Yilin Pan[1], *Bahman Mirheidari*[1], *Markus Reuber*[2,3], *Annalena Venneri*[4], *Daniel Blackburn*[4], *and Heidi Christensen*[1]

[1]Dept. of Computer Science, University of Sheffield, UK
[2]Academic Neurology Unit, University of Sheffield, UK
[3]Royal Hallamshire Hospital, UK
[4]Sheffield Institute for Translational Neuroscience, University of Sheffield, UK
{yilin.pan, b.mirheidari, heidi.christensen}@sheffield.ac.uk

Abstract

Speech and language based automatic dementia detection is of interest due to it being non-invasive, low-cost and potentially able to aid diagnosis accuracy. The collected data are mostly audio recordings of spoken language and these can be used directly for acoustic-based analysis. To extract linguistic-based information, an automatic speech recognition (ASR) system is used to generate transcriptions. However, the extraction of reliable acoustic features is difficult when the acoustic quality of the data is poor as is the case with DementiaBank, the largest opensource dataset for Alzheimer's Disease classification. In this paper, we explore how to improve the robustness of the acoustic feature extraction by using time alignment information and confidence scores from the ASR system to identify audio segments of good quality. In addition, we design rhythm-inspired features and combine them with acoustic features. By classifying the combined features with a bidirectional-LSTM attention network, the F-measure improves from 62.15% to 70.75% when only the high-quality segments are used. Finally, we apply the same approach to our previously proposed hierarchical-based network using linguistic-based features and show improvement from 74.37% to 77.25%. By combining the acoustic and linguistic systems, a state-of-the-art 78.34% F-measure is achieved on the DementiaBank task.

Index Terms: Dementia detection, automatic speech recognition, confidences score, acoustic feature

1. Introduction

Alzheimer's disease (AD) is the most common cause of neurodegenerative dementia, resulting in decline of memory, speech and other cognitive abilities. The number of people living with AD is increasing rapidly all around the world in the ageing society [1]. Spoken language, as one of the most important communication styles, can reveal an individual's cognitive ability. The research on patients' language and speech has revealed that language impoverishment and speech vagueness appear even at the early stage of dementia [2, 3].

From a linguistic point-of-view, people living with dementia present with symptoms at word-level such as having smaller vocabularies and Part of Speech (POS) misuse; sentence-level symptoms include incomplete sentences [4]. In addition to the linguistic impoverishment, people living with dementia also suffer from speech degeneration. The symptoms include, but are not limited to, phonological and articulatory impairments [5], high hesitation ratio, change in speech rhythm [6], and difference in fundamental frequency [7]. Automatic approaches to cognitive assessment rely on a combination of linguistic and acoustic information to detect symptoms more comprehensively [8]. Automatic system will typically work on audio recordings and all processing steps must be done automatically, including the acoustic feature extraction and the speech-to-text transcription by an Automatic Speech Recognition (ASR) system.

For acoustic feature extraction, the conventional approach is to calculate the features across the full audio recording. However, segments with high pause rates and/or high background noise in the recording may adversely affect the quality of the calculated features. Therefore, manual or automatic selection of sub-segments of the recordings with a *higher* acoustic quality may improve the performance of the extracted features [9, 10]. Confidence scores have been shown to improve the reliability of the automatic transcripts [11, 12] lending evidence to the benefit of relying more on high quality speech segments and transcribed words selection for automatic cognitive assessment.

Another output from the ASR decoding process is the time alignment, which can also be used for acoustic feature extraction. Rhythm is a speech property to do with the temporal organization of sounds [13]. It can be partially described by related statistical parameters such as speech unit duration, and the number and duration of pauses. It is known that speech rhythm varies significantly between healthy controls (HC) and people with AD [6, 13–17]. To extract the rhythmic parameters, manually identifying the word location is both time consuming and error prone, especially when done at scale. ASR information such as word alignment and confidence scores have been used for automatic rhythm feature construction [16, 18–20].

In this paper, DementiaBank, the largest open-source datasets for dementia classification [21], is used. The contributions of this paper are as follows: (1) A set of new rhythm-inspired features is proposed by utilizing the ASR decoding information (confidence scores and alignment). (2) we show that using ASR decoding information for selecting high-quality speech segments for robust acoustic feature extraction improves results. Further improvements are seen when combining with the rhythm-inspired features. (3) Applying the proposed technique to our linguistic-based system proposed in [22] also improves that system. (4) Finally, combining our acoustic and linguistic systems achieves a further improvement. To the best of our knowledge, the resulting F-measure of 78.34% is the highest achieved for the automatic-based DementiaBank task.

In the remainder of this paper, Section 2 presents the background and related work, Section 3 presents the proposed methodology, Sections 4 and 5 describe the experimental setup and results, and finally, the conclusions are given in Section 6.

http://dx.doi.org/10.21437/Interspeech.2020-2698

2. Background

DementiaBank, recorded between 1983 and 1988, is the most widely used opensource dataset for dementia detection. It contains acoustic recordings of people describing the *cookie theft* picture. The high background noise, variations in audio quality and the limited number of recordings increase the difficulties of speech analytics-based research investigating acoustic cues. Developing approaches that are robust to such challenges are essential for automatic systems to eventually make the leap from a lab setting to being reliable *in-the-wild*.

Electroencephalography (EEG) recordings from AD and HC shows a close link between the average duration of speech units and the frequency range of cortical oscillations that can be influenced by cognitive decline [13]. The change of speech and pause duration is constructed into rhythm features for AD detection. In addition to speech rhythm [15], duration features [13], speech fluency [14], semantic fluency [16], or pause location [6, 17] are also adopted to describe similar symptom in the literature. While modelling the rhythm characters, most of the research regards syllables or words equally without taking their quality into consideration. However, some symptoms of speech articulation disorders, like phonological errors or unclear pronunciations can also be found in AD [5]. Rhythm features, used for describing the organization of the sound, can be more robust if the adverse affect of the speech quality handled better.

In an ASR system, the acoustic segments tend to be difficult to recognize if the word articulation is 'blurred' or there is a lot of background noise, resulting in a low confidence score for the recognized word. In this paper, we propose to use the confidence score as a proxy measure for the quality of the spoken segments and the reliability of the recognized words. Specifically, a set of more informative and robust rhythm features is designed by utilizing word confidence scores for the speech pattern description. Additionally, the higher quality spoken segments and transcribed words are identified for extracting acoustic and linguistic information for the automatic system.

The extracted features can be classified by either the traditional linear classifiers, like Support Vector Machine (SVM) and Logistic Regression (LR), or neural networks (NNs), like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). Compared with the traditional linear classifiers, the NNs are generally more powerful on feature representation and classification. Among them, RNNs are good at capturing the temporal evolution of input signals and model the sequence information [23], which is suitable for the acoustic and rhythm feature modelling. The attention mechanism has lately been used in different fields and achieved a great deal of success [22,24]. The main idea behind the attention mechanism is applying a higher attention weight to the more critical parts of the input for classification. In this paper, to explore the time sequence information embedded in the acoustic and rhythm features, a bidirectional-LSTM (bi-LSTM) with attention mechanism is applied for classification.

3. Methodology

3.1. Data Analysis

Before describing the classification results, it is worth analysing the information output from the ASR decoding of DementiaBank and how this might vary for the AD and HC group. The ASR system (more information in Section 4.2) provides word identities, confidence scores, and word alignments. This information is used to calculate a number of parameters listed in Table 1. As shown, both the word and pause duration is longer on average in the AD group than in HC, while the number of words per transcript is lower in AD than in HC, which is consistent with the analysis in [6]. At the same time, the confidence score is higher for the HC group than the AD group. The discrepancy of these parameters indicate that using ASR decoding output might be informative for AD/HC classification.

Table 1: *The parameters analysis for the recordings and automatic transcripts from HC and AD of DementiaBank dataset.*

parameters (mean&var)	**HC**	**AD**
word duration (s)	.535 (0.984)	.642 (1.632)
pause duration (s)	.545 (1.091)	.663 (1.835)
#words/transcript	97 (3193)	84 (2476)
confidence score	.916 (0.029)	.882 (0.038)

A part of the waveform from the DementiaBank acoustic recordings is plotted in Figure 1, together with its corresponding manual and automatic transcripts. It is clear that the segment contains a lot of noise and has a long pause. Acoustic features are often extracted across the full signal (speech+pause), although some will try to identify and exclude the pause part. We propose to additionally exclude the parts of the speech with a low confidence score. On the figure, comparing the manual to the ASR transcript it can be seen that the word *I* has been mis-recognised as *let's*, and this word also has an associated low confidence score. Our approach would exclude this part of the speech as well (using the word alignments and identified by thresholding the confidence score).

Specifically, to get the high confidence/quality segments, the pause between two high confidence words is neglected if the duration is shorted than 0.1s, like the pause between *what* and *is*, and between *she* and *doing*. Finally, two audio segments and their corresponding transcribed words such as *see the mother* and *what's she doing* are selected for further processing. To measure the rhythm information embedded in the recordings, the start time, end time and the number of words for the high confidence segments are recorded as the three-dimensional rhythm features for the corresponding segment.

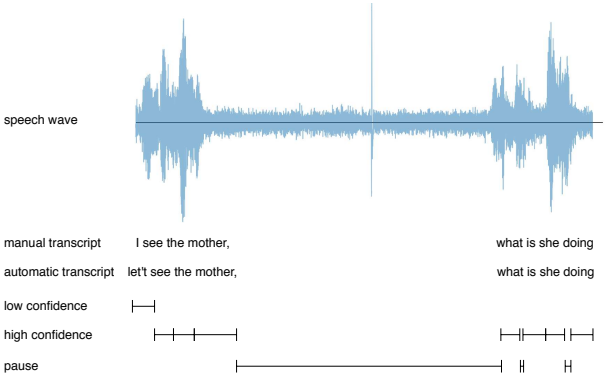

Figure 1: *Example of the high quality acoustic segments and automatic transcript selection; conf is set equal to* 0.95.

To represent the acoustic information, IS10, which achieved the best result on DementiaBank reported in [10], is adopted as the frame-level feature set. Then, the suprasegmental acoustic feature vector for each segment are generated by averaging the frame-level features over the high confidence segment.

3.2. Acoustic Features Classification

To classify the acoustic features of each recording composed into the suprasegmental feature vectors, a bi-LSTM with an attention mechanism is built as shown in Figure 2. The bi-LSTM contains two sub-networks for the forward and backward sequence information modelling. The output of the i_{th} segment h_i is represented by an element-wise sum on the outputs of the two sub-networks, and works as the input of the attention layer.

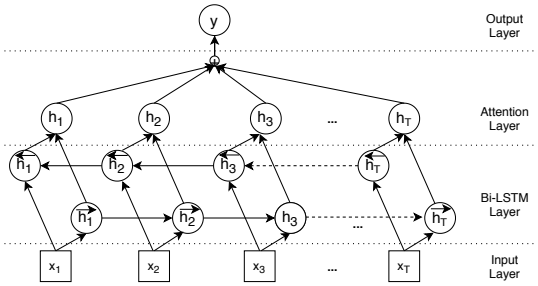

Figure 2: *Bidirectional LSTM model with attention*

The attention function is defined as follows:

$$\begin{aligned} u_i &= tanh(Wh_i + b) \\ \alpha_i &= \frac{exp(u_i^T u)}{\sum_i exp(u_i^T u)} \\ y &= \sum_i \alpha_i h_i. \end{aligned} \tag{1}$$

where u_i can be regarded as the hidden representation of h_i through a one-layer MLP. The importance of each hidden representation is measured by the normalized similarity between u_i and u. The vector u can be regarded as a high-level representation of the fixed query "what is the important information in the fixed input" [25]. It is randomly initialized and jointly learned during the training process. The final recording-level feature vector y is classified on a dense layer with a $sigmoid$ function.

4. Experiment Setup

4.1. Dataset

Our experimental task is the binary classification based on the DementiaBank dataset. Originally, there were 551 recordings from 293 speakers, but the diagnostic class for some of the participants changed during their follow-up (mostly for people with mild cognitive impairment (MCI) converting to AD). After removing them, 222 samples from 89 HC and 255 from 168 AD were selected. To evaluate our proposed method, speaker-independent 10-fold cross-validation (CV) was used. 8 folds were used for training (training set), 1 fold for evaluation (development set) and 1 fold for testing (test set). More information about the selected recordings and the 10-fold split were described in the paper [22] and the lists are available in GitHub[1].

4.2. Automatic Speech Recognition

To train the ASR, a 10-fold CV approach was also used with 9 folds of DementiaBank used for training and 1 for testing. The Kaldi's Librispeech [26] recipe was followed to train a base Time delay neural network (TDNN) acoustic model. Then using the transfer learning technique proposed by [27], 'transferring all layers', the acoustic model was adapted to the data in each fold (we followed a similar approach to [28] using only one epoch of training to get the best results). The language models were trained using the four-grams models gained from the transcripts in each fold interpolated with the four-gram of the Librispeech data set. To boost both the acoustic and language models, we added an extra dataset to the training set in each fold (the Hallamshire dataset [29]; 64 hours conversational recordings between doctors and patients). We achieved an average 32.3% WER for the ASRs, compared with 41.6% in [22].

4.3. Pre-Processing of Audio and ASR Transcripts

As in [22], though automatically adding punctuation can decrease the performance compared with manually punctuation, it can benefit the results compared with using the ASR transcripts directly without considering the sentence boundary. To add punctuation in the ASR transcripts, the toolkit shared in github[2] was used and more information can be found in [30]. After punctuation, only the words with confidence higher than the *conf* threshold were left in the transcripts. For the acoustic features, the 76 dimensional IS10 frame-level features were extracted with the OpenSMILE toolkit [31] from the selected segments, before calculating the mean vector across time. The 3-dimension rhythm feature was extracted from each segment.

4.4. Acoustic Classifier Configuration

For each fold, the bi-LSTM attention system described in Section 3.2 was trained with a 8-fold training set on a fixed number of epochs (20) and evaluated on the 1-fold development set at each epoch. The best model was selected based on the F-measure of the development set. All the results reported in our experiment were averaged across the 10-fold CV of the test set. The bi-LSTM units were set to 50 and attention layer dimension was 10. The batch size was set to 20. The maximum number of segments in each recording was set as 50 and zero-padding was used for recordings with fewer than 50 segments. Dropout with a rate of 0.5 was applied after the bi-LSTM and attention layer to avoid over-fitting. The network was optimized using *Adam* [32] with L2 regularization ($\lambda = e^{-6}$). For fairness a constant random seed was used. In the final combined system, the fusion was achieved by concatenating the output of the attention layer with the linguistic document-level representation trained by the second attention layer described in [22]. For the combination system, apart from using a different number of epochs (30) and dropout rate (0.3), all other parameters were kept unmodified as in the hierarchical attention network [22] and the acoustic system described in Section 4.4.

5. Result

5.1. Analysis of Rhythm-related Features

To analyse the relationship between the designed rhythm features at different thresholds and the diagnostic class, Pearson's correlation value (r-value) and p-value are adopted as the criteria for the analysis shown in Table 2. The conclusions are summarized as follow:

1. The three-dimensional rhythm features (duration of included segments, duration of excluded segments, and num-

[1]https://github.com/YilinSpeechandNLP/Automatic-Hierarchical-Attention-Neural-Network-for-Detecting-AD

[2]https://github.com/ottokart/punctuator2

Table 2: *The Pearson's correlation between the rhythm features with diagnostic class at different threshold; all p-values are* < 0.01.

Rhythm feature	threshold=0	threshold=0.5	threshold=0.8	threshold=0.9	threshold=0.95	threshold=0.99
Include duration (in s)	0.075	0.063	0.045	0.044	0.040	0.037
Exclude duration (in s)	0.136	0.153	0.183	0.171	0.173	0.169
Number of words included	-0.045	-0.071	-0.114	-0.137	-0.151	-0.176

ber of words in included segments) are all related to the confidence threshold as per their r-value.

2. For higher thresholds, the correlation increases between the number of words in the included segments and the diagnostic class, indicating that this features becomes more informative for classification in high-quality segments.
3. Compared with the duration of the excluded segments, the duration of the included segment is not that informative for classification. This should inform the rhythm feature extraction approach. In particular, more attention should be given to duration of pauses and the less clear words than the duration of the clearly spoken words.
4. Although the three-dimensional rhythm-related features show a different correlation between the threshold and the diagnostic class, increasing the threshold can increase the correlation between their absolute value addition and the diagnostic class.

5.2. Acoustic based Results

For evaluating the influence of the confidence score threshold on the classification performance, the relationship between the threshold and F-measure for the development & test set is shown in Figure 3 for the combination of IS10 and rhythm-related features and for IS10 only.

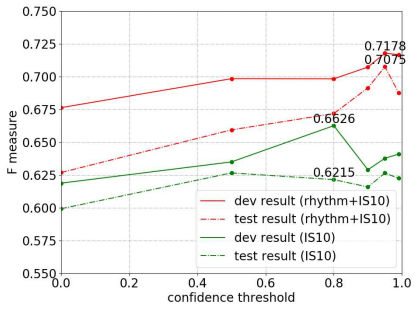

Figure 3: *The relationship between F-measure and word confidence threshold on evaluation set and test set.*

The conclusions are summarized as follows:

1. By comparing the results from the same feature sets, we find that increasing the threshold within a certain range can benefit the classification. The best F-measures on the development set is achieved with a threshold set to 0.8 and 0.95 on IS10 and IS10+rhythm feature sets respectively.
2. Comparing the red and blue lines with the same threshold proves the efficiency of the proposed rhythm features. The best test result improves from 62.15% to 70.75% after the three-dimensional rhythm-related features are included.

The results in Figure 3 show the benefit of the segment-selection approach on the acoustic features extraction. We further returned to our linguistic-based system in [22] to explore whether that system would also benefit. In the following experiment, the threshold is fixed at 0.95.

Table 3: *The result of linguistic-based system with/without acoustic-based system.*

Linguistic	Acoustic	F-measure (dev)	F-measure (test)
Original transcripts	-	78.92%	75.55%
High conf transcripts	-	84.75%	77.25%
High conf transcripts	IS10+rhythm	84.73%	78.34%

The transcripts produced by the ASR system are tested by the hierarchical system proposed in [22]. The result is shown in Table 3. Firstly, the system was rerun using the improved ASR system which saw the performance increase by 1.18% (from 74.37% [22] to 75.55%). After applying the segment selection approach proposed in this paper, the system using these transcripts with only high confidence words achieved an F-measure of 77.25%. The result demonstrates that low confidence words can decrease the quality of the transcripts for classification, and that is it better to ignore such segments. The last row of Table 3 reports the performance of the result of combining the two systems (acoustic and linguistic). After fusion, the system achieves a state-of-the-art result of 78.34%. Though better results were reported in [8, 33], those systems were not fully automatic. To the best of our knowledge, only the 67.21% F-measure (weighted average) and 62% F-measure (weighted average) reported in [9, 34] were achieved without any manual information included.

6. Conclusions

In this paper, a three-dimensional rhythm-related feature set was designed using the ASR decoding (confidence scores and alignment) information. Increasing the confidence score threshold within some range for segment selection was proven to be beneficial for the performance of the acoustic features extracted from the selected high-quality segments. Combining the designed rhythm features with acoustic features further improved the performance of the acoustic system. Finally, a state-of-the-art automatic system for the DementiaBank dataset was composed by utilizing the acoustic information and linguistic information selected by the confidence score threshold. In the future, the knowledge learned on DementiaBank is expected to be applied to other datasets for more practical applications.

7. Acknowledgements

This work is supported under the European Union's H2020 Marie Skłodowska-Curie programme TAPAS (Training Network for PAthological Speech processing; Grant Agreement No. 766287).

8. References

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

[2] S. Ahmed, A.-M. F. Haigh, C. A. de Jager, and P. Garrard, "Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease," *Brain*, vol. 136, no. 12, pp. 3727–3737, 2013.

[3] G. Szatloczki, I. Hoffmann, V. Vincze, J. Kalman, and M. Pakaski, "Speaking in alzheimer's disease, is that an early sign? importance of changes in language abilities in Alzheimer's disease," *Frontiers in aging neuroscience*, vol. 7, p. 195, 2015.

[4] R. S. Bucks, S. Singh, J. M. Cuerden, and G. K. Wilcock, "Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance," *Aphasiology*, vol. 14, no. 1, pp. 71–91, 2000.

[5] K. Croot, J. R. Hodges, J. Xuereb, and K. Patterson, "Phonological and articulatory impairment in Alzheimer's disease: a case series," *Brain and language*, vol. 75, no. 2, pp. 277–309, 2000.

[6] A. Pistono, M. Jucla, E. J. Barbeau, L. Saint-Aubert, B. Lemesle, B. Calvet, B. Koepke, M. Puel, and J. Pariente, "Pauses during autobiographical discourse reflect episodic memory processes in early Alzheimer's disease," *Journal of Alzheimer's Disease*, vol. 50, no. 3, pp. 687–698, 2016.

[7] A. Venneri, K. E. Forbes-Mckay, and M. F. Shanks, "Impoverishment of spontaneous language and the prediction of Alzheimer's disease," *Brain*, vol. 128, no. 4, pp. E27–E27, 2005.

[8] 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.

[9] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The ADReSS challenge," *arXiv preprint arXiv:2004.06833*, 2020.

[10] 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.

[11] A. Coucke, A. Saade, A. Ball, T. Bluche, A. Caulier, D. Leroy, C. Doumouro, T. Gisselbrecht, F. Caltagirone, T. Lavril *et al.*, "Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces," *arXiv preprint arXiv:1805.10190*, 2018.

[12] D. Yu, S. Wang, J. Li, and L. Deng, "Word confidence calibration using a maximum entropy model with constraints on confidence and word distributions," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 4446–4449.

[13] F. Martínez-Sánchez, J. J. Meilán, J. A. Vera-Ferrandiz, J. Carro, I. M. Pujante-Valverde, O. Ivanova, and N. Carcavilla, "Speech rhythm alterations in spanish-speaking individuals with Alzheimer's disease," *Aging, Neuropsychology, and Cognition*, vol. 24, no. 4, pp. 418–434, 2017.

[14] S. Ash, C. McMillan, R. G. Gross, P. Cook, D. Gunawardena, B. Morgan, A. Boller, A. Siderowf, and M. Grossman, "Impairments of speech fluency in Lewy body spectrum disorder," *Brain and language*, vol. 120, no. 3, pp. 290–302, 2012.

[15] S. Skodda and U. Schlegel, "Speech rate and rhythm in Parkinson's disease," *Movement disorders: official journal of the Movement Disorder Society*, vol. 23, no. 7, pp. 985–992, 2008.

[16] A. Satt, R. Hoory, A. König, P. Aalten, and P. H. Robert, "Speech-based automatic and robust detection of very early dementia," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[17] G. Angelopoulou, D. Kasselimis, G. Makrydakis, M. Varkanitsa, P. Roussos, D. Goutsos, I. Evdokimidis, and C. Potagas, "Silent pauses in aphasia," *Neuropsychologia*, vol. 114, pp. 41–49, 2018.

[18] W. Jarrold, B. Peintner, D. Wilkins, D. Vergryi, C. Richey, M. L. Gorno-Tempini, and J. Ogar, "Aided diagnosis of dementia type through computer-based analysis of spontaneous speech," in *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2014, pp. 27–37.

[19] L. Tóth, G. Gosztolya, V. Vincze, I. Hoffmann, G. Szatlóczki, E. Biró, F. Zsura, M. Pákáski, and J. Kálmán, "Automatic detection of mild cognitive impairment from spontaneous speech using ASR," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[20] A. Satt, A. Sorin, O. Toledo-Ronen, O. Barkan, I. Kompatsiaris, A. Kokonozi, and M. Tsolaki, "Evaluation of speech-based protocol for detection of early-stage dementia," in *Interspeech*, 2013, pp. 1692–1696.

[21] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle, "The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis," *Archives of Neurology*, vol. 51, no. 6, pp. 585–594, 1994.

[22] Y. Pan, B. Mirheidari, M. Reuber, A. Venneri, D. Blackburn, and H. Christensen, "Automatic hierarchical attention neural network for detecting AD," *Proc. Interspeech 2019*, pp. 4105–4109, 2019.

[23] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[24] M.-T. Luong, H. Pham, and C. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[25] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 1480–1489.

[26] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," IEEE Signal Processing Society, Tech. Rep., 2011.

[27] V. Manohar, D. Povey, and S. Khudanpur, "Jhu kaldi system for Arabic MGB-3 ASR challenge using diarization, audio-transcript alignment and transfer learning," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 346–352.

[28] B. Mirheidari, D. Blackburn, R. O'Malley, A. Venneri, T. Walker, A. Venneri, M. Reuber, and H. Christensen, "Improving cognitive impairment classification by generative neural network-based feature augmentation," *Interspeech*, 2020, submitted to Interspeech 2020.

[29] B. Mirheidari, D. Blackburn, R. O'Malley, T. Walker, A. Venneri, M. Reuber, and H. Christensen, "Computational cognitive assessment: Investigating the use of an intelligent virtual agent for the detection of early signs of dementia," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2732–2736.

[30] O. Tilk and T. Alumäe, "Bidirectional recurrent neural network with attention mechanism for punctuation restoration." in *Interspeech*, 2016, pp. 3047–3051.

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

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[33] M. Yancheva and F. Rudzicz, "Vector-space topic models for detecting Alzheimer's disease," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 2337–2346.

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