AN EXPLORATION OF LOG-MEL SPECTROGRAM AND MFCC FEATURES FOR ALZHEIMER'S DEMENTIA RECOGNITION FROM SPONTANEOUS SPEECH

Amit Meghanani, Anoop C. S., A. G. Ramakrishnan

Medical Intelligence and Language Engineering Laboratory
Indian Institute of Science, Bengaluru, India- 560012

ABSTRACT

In this work, we explore the effectiveness of log-Mel spectrogram and MFCC features for Alzheimer's dementia (AD) recognition on ADReSS challenge dataset. We use three different deep neural networks (DNN) for AD recognition and mini-mental state examination (MMSE) score prediction: (i) convolutional neural network followed by a long-short term memory network (CNN-LSTM), (ii) pre-trained ResNet18 network followed by LSTM (ResNet-LSTM), and (iii) pyramidal bidirectional LSTM followed by a CNN (pBLSTM-CNN). CNN-LSTM achieves an accuracy of 64.58% with MFCC features and ResNet-LSTM achieves an accuracy of 62.5% using log-Mel spectrograms. pBLSTM-CNN and ResNet-LSTM models achieve root mean square errors (RMSE) of 5.9 and 5.98 in the MMSE score prediction, using the log-Mel spectrograms. Our results beat the baseline accuracy (62.5%) and RMSE (6.14) reported for acoustic features on ADReSS challenge dataset. The results suggest that log-Mel spectrograms and MFCCs are effective features for AD recognition problem when used with DNN models.

Index Terms— log-Mel spectrogram, MFCC, transfer learning, Alzheimer, dementia, MMSE, CNN, LSTM, ResNet18.

1. INTRODUCTION

Dementia is a generic term for loss of memory and other thinking abilities that are severe enough to interfere with one's ability to deal with the activities of daily life. Alzheimer's disease is the most common form of dementia. The impact of dementia is not only restricted to the patients, but also to their carers, families, and society at large. According to the World Health Organization, there are around 50 million people affected by dementia worldwide, and there are nearly 10 million new cases every year [1]. 90% of them are aged over 65 years. Since there is no cure for the disease, early detection is critical to delay its progress and ensure better quality of life for the patients. Hence it is of paramount importance to investigate cost-effective and scalable methods for early detection of dementia.

Though the most distinguishable symptom of Alzheimer's dementia (AD) is memory deterioration, speech and language impairments are also common [2]. Speech is being used for the diagnosis of cognitive impairments/mental disorders like depression [3], [4], [5]. In recent years, many methods based on signal processing, machine learning, and natural language processing have been proposed for the task of AD classification using the audio data and/or the transcriptions. Fraser et al. [2] use both linguistic and acoustic information to train a machine learning based classifier. Mirheidari et al. [6] employ basic turn-taking statistics for dementia detection. Luz [7] uses a Bayesian classifier operating on patterns of vocalizations and other paralinguistic speech features, not relying on the transcriptions. Mirheidari et al. [8] make use of word vector representations of the individual words in the automatic speech transcription, for the classification task. They also employ CNN-LSTM networks on GloVe [9] word vectors for sequence classification of fixed length text taken from the input text using a sliding window. Chien et al. [10] use a token sequence of syllables to train a convolutional recurrent neural network for Alzheimer's disease assessment. Haider et al. [11] assess several acoustic features via extended Geneva minimalistic acoustic parameter set (eGeMAPS) [12], emobase, the ComParE 2013 [13], and multi-resolution cochleagram (MRCG) features [14] for the AD classification problem. They also propose a new active data representation method for feature extraction. Among the acoustic feature based classifiers, Luz etal. [15] report the best accuracy of 62.5% in the AD/Non-AD classification task on the test set of ADReSS challenge dataset, using ComParE features and an LDA classifier. They also report an RMSE of 6.14 in the MMSE prediction, which is obtained using decision trees (DT) on MRCG features. We consider this as the baseline for our work.

Log-Mel spectrograms and MFCCs are being used extensively in deep learning frameworks for various tasks, such as emotion recognition [16][17], audio classification [18], detection of cognitive impairments/mental disorders like depression [4][5], and automatic speech recognition (ASR) [19]. However, they are not much explored for the evaluation of speech impairments associated with AD. In this work we evaluate the effectiveness of log-Mel spectrogram and MFCC features for the following tasks on ADReSS challenge dataset [15] with deep learning frameworks:

1. Binary classification of spontaneous speech samples into AD and non-AD classes.
2. Prediction of mini-mental state examination (MMSE) score, an indicative measure of cognitive impairment.

For a fair comparison, we compare our results only with acoustic feature based methods in the literature.

The rest of the paper is organized as follows. Section 2 describes the ADReSS challenge dataset. Sections 3 and 4 explain our approaches in AD classification and MMSE score regression tasks. Section 5 discusses the conclusions of our work.

2. ADRESS CHALLENGE DATASET

The ADReSS challenge dataset consists of speech recordings and transcripts of spoken picture descriptions elicited from participants. It is balanced for age and gender to minimize the risk of bias in the prediction tasks. Speech data segmented with voice activity detection (VAD) algorithms is also provided. The training set has data from 108 subjects, 54 each from AD and non-AD classes. Each class has 24 male and 30 female subjects with age between 50 and 80. Test set is also balanced with a total of 48 subjects. Each class in the test set has 11 male and 13 female subjects. The range of ages the is same as the training data.

3. AD CLASSIFICATION TASK

The AD classification task is a binary classification problem to distinguish between AD and non-AD subjects. The evaluation metric for this task are accuracy = $\frac{TN+TP}{N}$, precision $\pi = \frac{TP}{TP+FP}$, recall $\rho = \frac{TP}{TP+FN}$, and $F_1 = 2\frac{\pi\rho}{\pi+\rho}$, where N is the total number of subjects involved in the study, TP, FP, TN and FN are the number of true positives, false positives, true negatives and false negatives, respectively.

3.1. Feature extraction

All the speech samples are downsampled to 16 kHz. We use log-Mel spectrograms and MFCCs for our analysis. These features together with their delta and delta-delta, are extracted using librosa library [20].

3.1.1. Details of extraction of log-Mel spectrogram

We choose 224 Mel filter banks for generation of log-Mel spectrograms. The choice of 224 Mel filter banks allows us to use the pretrained CNN models like ResNet, down the line. We use a Hanning window of size 2048 samples ($\approx$ 128 ms) and hop-length (the number of samples between successive frames) of 512 samples ($\approx$ 32 ms). The number of points in the FFT computation is also 2048. The log-Mel spectrogram, delta, and delta-delta images are scaled down to the values between 0 and 1 using scikit-learn library [21]. CNN architectures at the front end necessitate fixed-size images at the input. To achieve this, the log-Mel spectrograms are divided temporally to non-overlapping segments of 224 frames each covering a duration of 7.264 seconds. Log-Mel spectrogram, delta, and delta-delta features are fed respectively to the channels 1, 2, and 3 of the CNN input layer. Thus, input to the CNN has a dimension of 3x224x224 (channels x height x width). This choice of this feature dimension is motivated by the fact that pretrained models like ResNet should have a minimum dimension of 3x224x224 at its input.

pBLSTM-CNN uses log Mel spectrogram features generated with a window size of 400 samples (25 ms) and hop-length of 160 samples (10 ms). 512-point FFT and 40 Mel-scale filter banks are used. The delta and delta-delta features are also used, resulting in an input dimension of 120.

3.1.2. Details of MFCC extraction

For MFCC, we use a Hanning window of size 480 samples (30 ms), hop- length of 160 samples (10 ms), and 512-point FFT. We compute 40 MFCCs along with their delta and delta-delta. The obtained MFCCs are divided temporally to non-overlapping segments of 300 frames ($\approx$ 3 s). MFCCs, delta, and delta-delta features are fed respectively to the channels 1, 2, and 3 of the CNN input layer. Thus, input to the CNN has a dimension of 3x40x300 (channels x height x width).

3.2. DNN architectures

3.2.1. CNN-LSTM

In speech-based applications such as ASR, the CNN has become an attractive model, as it can transform speech signals to feature maps as in computer vision applications [22]. CNN can exploit the local correlations of the speech signals in both time and frequency dimensions. CNN followed by LSTM networks [23] are widely used in speech domain to address tasks such as emotion recognition [17] and music classification [24]. Here, we use an end-to-end, fully trainable CNN-LSTM architecture to explore its capability in the AD classification task using log-Mel spectrogram and MFCC features. The architecture consists of a CNN for feature extraction from the input data, followed by LSTM cells to capture the sequential patterns in the features. The model is trained by backpropagating the error from the LSTM output through the LSTM cells to the layers in the CNN. Figure 1 shows the architecture of the CNN-LSTM model for log-Mel spectrograms as input features. Tables 1 and 2 list the details of the CNN and LSTM layers, respectively. The CNN consists of five convolution layers with 16, 32, 64, 128, and 256 filters respectively. The kernel size and stride at all convolution layers are 3x3 and

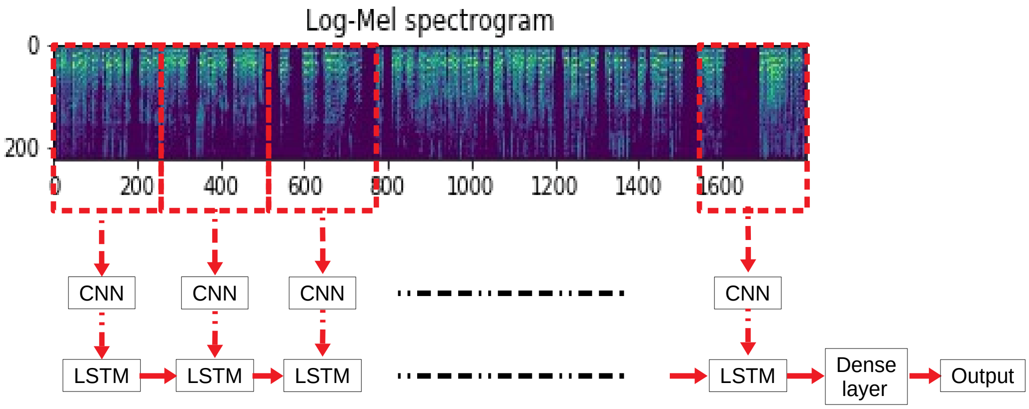

Fig. 1. Block diagram of log-Mel spectrogram based CNN-LSTM network.

Table 1. Model summary of CNN module in the CNN-LSTM network. Input to the CNN is log-Mel spectrogram image with dimension 3x224x224.

Layers	Input dim	Operations	Output dim
Conv1	3x224x224	Conv2d BatchNorm2d Max_Pool2d ReLU	16x111x111
Conv2	16x111x111	Conv2d BatchNorm2d Max_Pool2d ReLU Dropout	32x54x54
Conv3	32x54x54	Conv2d BatchNorm2d Max_Pool2d ReLU Dropout	64x26x26
Conv4	64x26x26	Conv2d BatchNorm2d Max_Pool2d ReLU Dropout	128x12x12
Conv5	128x12x12	Conv2d	256x10x10
Global Avg Pooling	256x10x10	AvgPool2d ReLU	256x1x1

Table 2. Model summary of LSTM module in the CNN-LSTM network. Input to the LSTM is [seq_len, input_dim], where seq_len is the length of the input sequence and input_dim is the dimension of the input feature vector. For CNN-LSTM, input_dim = 256 and for Resnet-LSTM, input_dim = 512.

Layers	Input dim	Output dim
Recurrent (LSTM)	[seq_len, input_dim]	64
Fully connected	64	2

1x1, respectively. The max-pooling layer has a kernel size of 2x2. Dropout probability is 0.1. LSTM accepts a sequence of 256-dimensional vectors as input and outputs the final hidden state of 64 dimensions.

For MFCCs, the parameters of input, output and convolution layers of the CNN are selected based on the input dimensions (3x40x300). The CNN consists of three convolution layers with 16, 32, and 64 filters respectively. The kernel size and stride at all convolution layers are 2x2 and 1x1, respectively. In this case, LSTM accepts a sequence of 64-dimensional vectors as input and outputs the final hidden state of 64 dimensions. All the other parameters remain the same.

3.2.2. *ResNet-LSTM model*

Here, the CNN is replaced by ResNet18 [25] (pre-trained on ImageNet) which can provide better generalization capacity and more expressive power, through its residual connections between layers. ResNet18 is used as a fixed feature extractor from the input log-Mel spectrogram images. Here, we freeze all the parameters of the ResNet18 network except the final layer. The output from the final layer of ResNet18 is fed as a sequence to the LSTM network. This setup is known as transfer learning, where the model trained for one particular problem is applied to another problem. Transfer learning has already been adopted successfully to improve the accuracy of speech-based emotion recognition systems which use Mel spectrogram as input features [26]. The output from the last layer of ResNet18 is a 512-dimensional feature vector. LSTM accepts a sequence of these 512-dimensional vectors as input. All other configurations of the LSTM model remain the same. We are not using MFCC features, as it has restrictions on input dimensions (3x224x224).

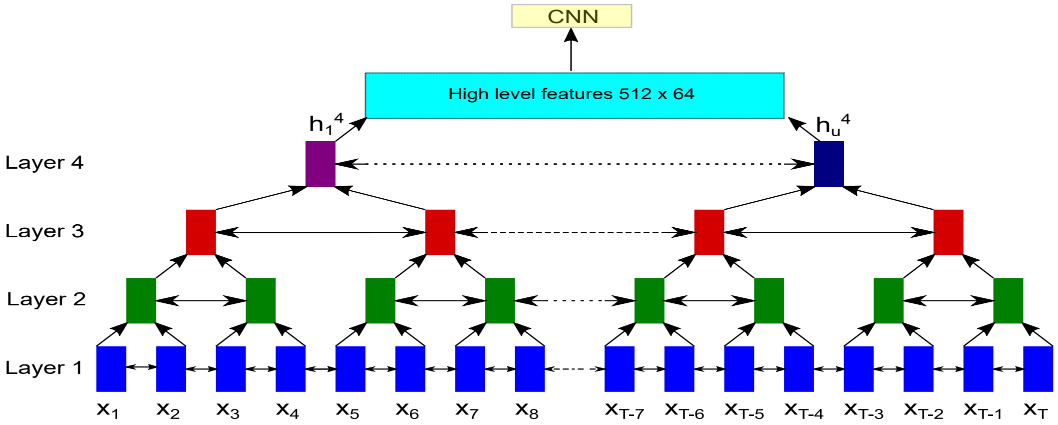

Fig. 2. High level feature extraction using pBLSTM

3.2.3. pBLSTM-CNN model

pBLSTMs are used in end-to-end speech recognition models like Listen, attend and spell [27]. In this approach, we use a pyramidal structure of bidirectional LSTM to convert the input features into high level features, which are later used by the CNN for classification. We use 4 layers of pyramidal bidirectional structure. At each layer, we combine the outputs of two successive time steps, before feeding it to the next layer. This reduces the number of time steps by a factor of 2 at each layer. This results in an overall compression by a factor of 16 at the output of the fourth layer of the pBLSTM.

$$x_t^l = [h_{2t}^{l-1},\ h_{2t+1}^{l-1}] \quad (1)$$

$$h_t^l = pBLSTM(h_{t-1}^l,\ x_t^l) \quad (2)$$

where x_t^l is the input to the layer l at time step t and h_t^l is the hidden state of the pBLSTM at layer l and time step t.

The high level features computed by the pBLSTM layers are limited to a maximum step size of 512. The hidden dimension of each pBLSTM layer is 32. Thus the input to the CNN architecture has dimensions 512x64. The CNN architecture employs 4 convolution layers, each followed by batch normalisation, max-pooling and ReLU operations. Each layer expands the depth of convolution by a factor of 4. The first 3 layers use kernels of 5x3, padding of (2,1) and stride of (2,1). The fourth layer has kernel size 3x3, padding (1,1) and stride (1,1). The output is fed to an average pooling layer of size 4 to obtain a 256 dimensional vector, which is fed to a fully connected layer to form the final output.

3.3. Training details

For the classification task, the DNN models are trained with a batch size of 16. An Adam optimizer with a learning rate of 0.001 and a step decay of 0.5 after every 25 epochs is used. The maximum number of epochs is set to 100. Early stopping is applied with a patience value of 20 to prevent the network from overfitting. Since AD classification is a two class problem, binary cross-entropy is used as the loss function.

Table 3. 5-fold cross-validation accuracy for AD classification task on the training subset of the ADReSS dataset.

Model	**log-Mel**	**MFCC**
CNN-LSTM	66.66%	59.13%
ResNet-LSTM	67.54%	–
pBLSTM-CNN	53.8%	61.68%

3.4. Results

3.4.1. 5-fold cross-validation

We have performed 5-fold cross-validation, to estimate the generalization error. The results of cross-validation on models trained using full audio data using log-Mel spectrogram and MFCC as the front end features are shown in Table 3. We are not using ResNet-LSTM with MFCC features, as it has restrictions on input dimensions (3x224x224). This is not possible with MFCCs as input features.

3.4.2. Bootstrap aggregation of DNN models

We have used bootstrap aggregation of models (known as bagging [28]) to predict the final labels for test samples. Bootstrap aggregation is an ensemble technique to improve the stability and accuracy of machine learning models. It combines the prediction from multiple models. It also reduces variance and helps to avoid overfitting.

Using the training set of size 108 (54 AD and 54 non-AD samples), we generate 21 new training sets, each of size 108, by sampling from the training set uniformly, with replacement. Sampling with replacement results in some samples

Table 4. Test results for AD classification task using bootstrap aggregation of 21 classifiers trained separately for CNN-LSTM, ResNet-LSTM and pBLSTM-CNN along with the baseline [15] results.

Model	Features	Class	Precision	Recall	F1 Score	Accuracy
CNN-LSTM	log-Mel	Non-AD	0.57	0.62	0.60	58.33%
		AD	0.59	0.54	0.56	
CNN-LSTM	MFCC	Non-AD	0.59	0.92	0.72	**64.58%**
		AD	0.82	0.38	0.51	
ResNet-LSTM	log-Mel	Non-AD	0.62	0.62	0.62	**62.50%**
		AD	0.62	0.62	0.62	
pBLSTM-CNN	log-Mel	Non-AD	0.53	0.42	0.47	52.08%
		AD	0.52	0.63	0.57	
pBLSTM-CNN	MFCC	Non-AD	0.63	0.5	0.56	60.42%
		AD	0.59	0.71	0.64	
Baseline - LDA	ComParE	Non-AD	0.67	0.50	0.57	**62.50%**
		AD	0.60	0.75	0.67	

Table 5. Type 1 and type 2 errors (out of 24 test samples each from the AD and non-AD classes) for AD classification task using bootstrap aggregation of 21 classifiers trained separately for CNN-LSTM, ResNet-LSTM and pBLSTM-CNN along with the baseline classifier.

Model	Features	Type 1 error (FP)	Type 2 error (FN)
CNN-LSTM	log-Mel	9	11
CNN-LSTM	MFCC	2	15
ResNet-LSTM	log-Mel	9	9
pBLSTM-CNN	log-Mel	14	9
pBLSTM-CNN	MFCC	12	7
Baseline - LDA	ComParE	12	6

being repeated, in each of the new training sets. These new training sets are known as bootstrap samples. 21 models are fitted using the above 21 bootstrap samples and the outputs are combined by a majority voting scheme for final classification. Examples not selected in a given bootstrap sample are used as the validation set to estimate the performance of the model. The performance of our CNN-LSTM, ResNet-LSTM and pBLSTM-CNN architectures on the test set, using bootstrap aggregation of 21 models, is tabulated in Table 4. The corresponding type 1 and type 2 errors are presented in Table 5. CNN-LSTM with MFCCs and ResNet-LSTM with log-Mel spectrograms achieve accuracies of 64.58% and 62.5%, respectively in the AD classification task. Among the acoustic feature based classifiers, our CNN-LSTM with MFCC outperforms other classifiers including the baseline (62.5%) [15], while ResNet-LSTM with log-Mel spectrogram, achieves the same accuracy as that of the baseline.

4. MMSE PREDICTION TASK

We address the problem of prediction of mini-mental state examination score by generating regression models using the CNN-LSTM and pBLSTM-CNN architectures and log Mel spectrogram features.

4.1. Details of the features used

Log-Mel spectrograms and MFCCs with their delta and delta-delta are used as the features for the MMSE prediction task also. The extraction of features has already been discussed in section 3.1.

4.2. DNN architectures

4.2.1. CNN-LSTM model

The same CNN-LSTM architecture described in section 3.2.1 is used for the regression task except for the changes in the output layer. The output layer is a fully connected layer with a linear activation function. The targets are normalized by the maximum MMSE score of 30, during training. The network is trained to minimize the mean squared error.

4.2.2. ResNet-LSTM model

The ResNet-LSTM network used for classification is modified with the output layer as linear and loss function as mean squared error (MSE). The targets are normalised by the maximum MMSE score of 30.

4.2.3. pBLSTM-CNN model

Here also the architecture is the same as that for classification except for the output layer, which is linear with mean squared error (MSE) loss function.

4.3. Training details

For CNN-LSTM and ResNet-LSTM, Adam optimizer is used with a learning rate of 0.001 and step decay of 0.5 every 10 steps. Early stopping with a patience value of 20 is applied to prevent the network from overfitting during training. For pBLSTM-CNN, a learning rate of 0.0001 is used and the maximum epochs is limited to 35.

4.4. Results

4.4.1. Bootstrap aggregation of DNN models

The final regression results on the test set are obtained with the bagging of 21 models by averaging the outputs. The root mean squared errors (RMSE) on the test set are shown in Table 6. The log-Mel spectrograms give RMSE of 5.90 and 5.98 on pBLSTM-CNN and ResNet-LSTM networks, respectively. The values suggest relative improvements of 2.6% and 3.9% over the baseline value of 6.14[15].

5. DISCUSSION AND CONCLUSIONS

In this work, we explore the usefulness of log-Mel spectrogram and MFCC features for the task of AD classification.

Table 6. RMSE of MMSE score predicted on the test set using bootstrap aggregation of 21 regressors trained separately for CNN-LSTM, ResNet-LSTM and pBLSTM-CNN along with the baseline [15] results.

Model	**Feature**	**RMSE**
CNN-LSTM	log-Mel	6.33
	MFCC	6.24
ResNet-LSTM	log-Mel	5.98
pBLSTM-CNN	log-Mel	**5.90**
	MFCC	6.71
Baseline (DT)	MRCG	6.14

We also explore the DNN based architectures for AD detection problem, as most of the classifiers in the literature are based on SVM, random forests, or decision trees. Our CNN-LSTM with MFCC and ResNet-LSTM with log-Mel spectrogram achieve accuracies of 64.58% and 62.5%, respectively in the AD classification task on the ADReSS challenge dataset. These results are similar or better than the baseline accuracy reported on the ADReSS challenge dataset using only the acoustic features (62.5% with ComParE features and an LDA classifier [15]). In the MMSE prediction task, the log-Mel spectrograms give an RMSE of 5.9 and 5.98, on pBLSTM-CNN and ResNet-LSTM models respectively, indicating relative improvements of 3.9% and 2.6% over the baseline RMSE of 6.14 [15]. Based on the results, we can conclude that log-Mel spectrograms and MFCCs are effective for the AD detection problem when they are incorporated with DNN models. Transfer learning based ResNet-LSTM performs well in both AD classification and MMSE score prediction. The application of ResNet18 prior to the LSTM seems to capture the spatial artifacts in the log-Mel spectrogram, relevant to the discrimination of AD. This architecture also helps to reduce the number of time steps in the LSTM, thereby improving the learning capability of the LSTM network. pBLSTM-CNN model performs well in the MMSE prediction task. The pyramidal structure in pBLSTM helps to reduce the number of time steps, allowing the CNN layers to focus on the artifacts relevant to MMSE prediction.

6. FUTURE WORK

End-to-end deep learning models enable us to do away with the hand-crafted features like MFCCs and Mel-spectrogram, and learn the relevant features directly from the raw waveform. They have been explored in speech recognition [29], [30], [31], speaker verification [32] and other speech related applications [33]. As a future work, we plan to employ end-to-end models in learning customised features for AD-classification task with the raw speech waveform as input. Another aspect we would like to focus upon is the explainability of the DNN models used in AD detection. Though DNN models are promising, it is inherently difficult to understand which aspects of the input feature is responsible for the decisions of the model. Since the results reported in this work use bagging of classifiers, it is a bit more challenging to trace the errors and the most useful features.

7. REFERENCES

[1] World Health Organization, "Mental health action plan 2013–2020," *WHO Library Cataloguing-in-Publication DataLibrary Cataloguing-in-Publication Data*, pp. 1–44, Jul. 2013.

[2] Kathleen Fraser, Jed Meltzer, and Frank Rudzicz, "Linguistic features identify Alzheimer's disease in narrative speech," *Journal of Alzheimer's disease : JAD*, vol. 49, Oct. 2015.

[3] Yasin Özkanca, Miraç Göksu Öztürk, Merve Ekmekci, David Atkins, Cenk Demiroglu, and Reza Hosseini Ghomi, "Depression screening from voice samples of patients affected by parkinson's disease," *Digital Biomarkers*, vol. 3, pp. 72–82, 06 2019.

[4] Lang He and Cui Cao, "Automated depression analysis using convolutional neural networks from speech," *Journal of Biomedical Informatics*, vol. 83, pp. 103 – 111, 2018.

[5] Emna Rejaibi, Ali Komaty, Fabrice Mériaudeau, S. Agrebi, and A. Othmani, "MFCC-based recurrent neural network for automatic clinical depression recognition and assessment from speech," *ArXiv*, vol. abs/1909.07208, 2019.

[6] Bahman Mirheidari, Daniel Blackburn, Markus Reuber, Traci Walker, and Heidi Christensen, "Diagnosing people with dementia using automatic conversation analysis," in *Proceedings of INTERSPEECH 2016*, San Francisco, CA, Sept. 2016, pp. 1220–1224.

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

[8] Bahman Mirheidari, Daniel Blackburn, Traci Walker, Annalena Venneri, Markus Reuber, and Heidi Christensen, "Detecting signs of dementia using word vector representations," in *Proceedings of INTERSPEECH 2018*, Sept. 2018, pp. 1893–1897.

[9] Jeffrey Pennington, Richard Socher, and Christoper Manning, "Glove: Global vectors for word representation," in *EMNLP*, 01 2014, vol. 14, pp. 1532–1543.

[10] Yi Wei Chien, Sheng Yi Hong, Wen Ting Cheah, Li Hung Yao, Yu Ling Chang, and Li Chen Fu, "An automatic assessment system for Alzheimer's disease based on speech using feature sequence generator and recurrent neural network," *Nature Scientific Report*, vol. 9, no. 19597, Dec. 2019.

[11] 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.

[12] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, 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.

[13] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," *MM 2013 - Proceedings of the 2013 ACM Multimedia Conference*, pp. 835–838, Oct. 2013.

[14] 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.

[15] Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The ADReSS Challenge," in *Proceedings of INTERSPEECH 2020*, Shanghai, China, 2020.

[16] H. Meng, T. Yan, F. Yuan, and H. Wei, "Speech emotion recognition from 3D log-Mel spectrograms with deep learning network," *IEEE Access*, vol. 7, pp. 125868–125881, 2019.

[17] Caroline Etienne, Guillaume Fidanza, Andrei Petrovskii, Laurence Devillers, and Benoit Schmauch, "CNN+LSTM architecture for speech emotion recognition with data augmentation," Sept. 2018, pp. 21–25.

[18] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, "Large-scale weakly supervised audio classification using gated convolutional neural network," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 121–125.

[19] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *INTERSPEECH*, 2019.

[20] Brian McFee, Colin Raffel, Dawen Liang, Daniel Patrick Whittlesey Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto, "Librosa: Audio and music signal analysis in Python," 2015.

[21] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al., "Scikit-learn: Machine learning in Python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[22] Yann LeCun and Yoshua Bengio, *Convolutional Networks for Images, Speech, and Time Series*, p. 255–258, MIT Press, Cambridge, MA, USA, 1998.

[23] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 677–691, 2017.

[24] Keunwoo Choi, György Fazekas, Mark B. Sandler, and Kyunghyun Cho, "Convolutional recurrent neural networks for music classification," *CoRR*, vol. abs/1609.04243, 2016.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.

[26] Yuanyuan Zhang, Jun Du, Zi-Rui Wang, and Jianshu Zhang, "Attention based fully convolutional network for speech emotion recognition," *CoRR*, vol. abs/1806.01506, 2018.

[27] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4960–4964.

[28] Leo Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

[29] T. Sainath, Ron J. Weiss, A. Senior, K. Wilson, and Oriol Vinyals, "Learning the speech front-end with raw waveform cldnns," in *INTERSPEECH*, 2015.

[30] Neil Zeghidour, Nicolas Usunier, Gabriel Synnaeve, Ronan Collobert, and Emmanuel Dupoux, "End-to-end speech recognition from the raw waveform," in *Proc. Interspeech 2018*, 2018, pp. 781–785.

[31] A. Madhavaraj and A. G. Ramakrishnan, "Scattering transform inspired filterbank learning from raw speech for better acoustic modeling," in *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)*, 2019, pp. 1154–1158.

[32] H. Muckenhirn, M. Magimai.-Doss, and S. Marcell, "Towards directly modeling raw speech signal for speaker verification using CNNs," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4884–4888.

[33] Juliette Millet and Neil Zeghidour, "Learning to detect dysarthria from raw speech," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5831–5835, 2019.