
Proceedings Paper:
Pan, Y., Mirheidari, B., Tu, Z. et al. (6 more authors) (2020) Acoustic feature extraction with interpretable deep neural network for neurodegenerative related disorder classification. In: Proceedings of Interspeech 2020. Interspeech 2020, 25-29 Oct 2020, Shanghai, China. International Speech Communication Association (ISCA) , pp. 4806-4810.

10.21437/Interspeech.2020-2684

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

Acoustic Feature Extraction with Interpretable Deep Neural Network for Neurodegenerative related Disorder Classification

Yilin Pan[1], *Bahman Mirheidari*[1], *Zehai Tu*[1], *Ronan O'Malley*[2], *Traci Walker*[3], *Annalena Venneri*[4], *Markus Reuber*[5], *Daniel Blackburn*[2], *Heidi Christensen*[1]

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

Abstract

Speech-based automatic approaches for detecting neurodegenerative disorders (ND) and mild cognitive impairment (MCI) have received more attention recently due to being non-invasive and potentially more sensitive than current pen-and-paper tests. The performance of such systems is highly dependent on the choice of features in the classification pipeline. In particular for acoustic features, arriving at a consensus for a best feature set has proven challenging. This paper explores using deep neural network for extracting features directly from the speech signal as a solution to this. Compared with hand-crafted features, more information is present in the raw waveform, but the feature extraction process becomes more complex and less interpretable which is often undesirable in medical domains. Using a SincNet as a first layer allows for some analysis of learned features. We propose and evaluate the Sinc-CLA (with SincNet, Convolutional, Long Short-Term Memory and Attention layers) as a task-driven acoustic feature extractor for classifying MCI, ND and healthy controls (HC). Experiments are carried out on an in-house dataset. Compared with the popular hand-crafted feature sets, the learned task-driven features achieve a superior classification accuracy. The filters of the SincNet is inspected and acoustic differences between HC, MCI and ND are found.
Index Terms:Neurodegenerative disorders, Mild cognitive impairment, feature interpretability, SincNet

1. Introduction

Neurodegenerative disorders (ND) are caused by slow progressive loss of neurons in the central nervous system leading to an irreversible selective loss of brain functions causing dementia. With an aging society, the number of people living with ND is increasing rapidly. Before being diagnosed as ND, people with early signs of cognitive decline often get diagnosed with Mild Cognitive Impairment (MCI). They exhibit symptoms worse than those expected from normal aging but not severe enough to be diagnosed as dementia [1]. About 10% to 15% of people living with MCI will *convert* into having Alzheimer's Disease (the most common type of dementia) per year [2]. Accurate and early detection of MCI and ND is of great importance.

Current clinical practice uses patient history and cognitive screening instruments plus structural brain imaging to exclude other causes - also known as a *rule-out* approach. The availability of expert neuropsychological testing is variable and is subject to long delays. Thus it is not feasible for wide-scale screening. In research settings and in some specialist centres some 'rule-in' diagnostic tests are used but these are either expensive and/or invasive, such as Positron Emission Tomography scanning for amyloid (Amyloid-PET) or analysis of the cerebrospinal fluid for biomarkers.

Even though memory impairment is the main symptom of MCI and ND, language and speech are also affected – even decades before diagnosis [3]. Recently, automatic approaches to analysing a person's speech and language have gained traction. Language-based analysis is mostly carried out on either the manual or automatic transcripts [4, 5], whereas speech-based analysis would normally be based on the acoustic signal [6–11]. In both cases, the performance of a typical classification pipeline is highly dependent on the quality of the front-end features. This paper focuses on finding better acoustic features. Conventional *hand-crafted* acoustic features can be classified into two classes: a group of general features, like MFCC [9] , F_0 [6], Jitter and Shimmer [7], and more specifically designed features informed by medical knowledge, like the features proposed in [10]. The general acoustic features contain information about voice quality, but cannot describe task-specific symptoms well often resulting in researcher opting to extract very long lists of features (often in the thousands) but still achieving unsatisfactory performance. On the other hand, the specially designed features require an exact translation from human's medical knowledge into mathematical expression, which can be challenging.

There are very few publicly available datasets for investigating cognitive decline, and most research is carried out on self-collected datasets which introduces a large variation in accents, background noise and the collecting device. As a result, feature sets found to be optimal for one dataset cannot necessarily display a stable performance on other datasets. Task-driven, learned features can be a better choice when the aim is generalisation. Neural networks (NNs) have proven their efficiency in various tasks as a front-end feature extractor [12, 13] compared with traditional hand-crafted acoustic features. However, most of the NNs appear as a black box, which means it is harder to analyse and interpret any learned representations which could have led to meaningful insights. In this paper we introduce the SincNet as a first NN layer in order to addressed this.

The contributions of this paper are as follows: (1) a feature extractor is constructed with a SincNet-fronted NN architecture for generating task-driven acoustic features; the performance on classifying MCI and ND from HCs (healthy controls) is much

http://dx.doi.org/10.21437/Interspeech.2020-2684

improved compared with the baseline feature sets. (2) an analysis of the SincNet reveals what information has been learned while training for classification. (3) to the best of our knowledge, this is the first study that explores the critical acoustic information for cognitive decline detection from a perspective of deep learning.

In the remainder of this paper, Section 2 presents the background. Section 3 presents the designed feature extractor. Sections 4 and 5 describe the experimental setup and results, and finally, the conclusions are given in Section 6.

2. Background

Feature extraction is crucial for the performance of a classification system. Depending on the task and dataset, hand-crafted features might not always be the best choice. For example, the Mel-scale filter bank designed to mimic auditory and physiological evidence of how humans perceive speech signals [14] is used broadly but cannot always guarantee to be the best filter bank for the target task. Compared with hand-crafted features, the raw wave includes more information. Extracting the target information directly from the raw waveform by NNs has been an active and promising area of research, especially for mainstream speech research fields like speech recognition [15], speaker recognition [16] and emotion recognition [17].

Convolutional neural networks (CNNs), deep neural networks (DNNs), and recurrent neural networks (RNNs) (long short-term memory (LSTM) and gated recurrent units (GRU)) are three of the most popular NN structures for speech processing applications. They have different advantages. CNNs have demonstrated their ability to extract robust and invariant representations when facing the typical frequency variations of acoustic recordings by applying local filters and pooling networks [18]. RNNs are good at capturing the temporal evolution of speech signals and model the sequence information [19]. In comparison, DNNs are generally used for mapping the features from one feature space into a more separable space. In [20], it was found that combining CNNs, LSTMs, and DNNs for speech processing in a unified architecture allowed for the exploitation of their complementary natures. The attention mechanism has lately been used in different fields and achieved a great deal of success [4, 21, 22]. The main idea behind the attention mechanism is to apply a higher attention weight to the more critical parts of the input for classification.

The first layer is always significant for the performance of the raw wave input system as it deals with the high-dimensional and noisy input [23]. Commonly used CNNs work as a task-specific finite impulse-response filterbank followed by a non-linearity [13]. A novel CNN structure named SincNet has been proposed. It benefits from having fewer parameters to learn. The filters are defined with a set of parametrized sinc functions and fewer paramaters need to be trained, making it more *interpretable* and the ability to converge faster [16]. These characteristics make it suitable for the first layer in our system.

This paper aims at building a system that can make use of the benefits of different kinds of networks for classifying ND and MCI from HCs. The dataset is a small-scale, self-collected dataset named IVA. It comprises of audio recordings of HCs and people living with MCI and ND as they interact with an Interactional Virtual Agent that asks them memory-probing questions (please, see Section 4.1 for more details about the data). In the system, a SincNet is applied as the first layer of our network followed by CNN (C), LSTM (L) and the attention mechanism (A); we refer to this network as Sinc-CLA in the following. Our

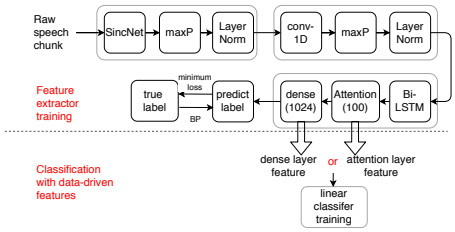


Figure 1: *The structure of the Sinc-CLA feature extractor.*

results show that the network-learned features can be more distinctive and informative compared to the INTERSPEECH 2010 Paralinguistic Challenge (IS10) feature set [24] as well as the ComParE 2013 feature set [25]. The structure of the Sinc-CLA system is illustrated in Figure 1.

3. Task-driven Feature Extraction

In this section, the process of task-driven feature extraction is described in details. The first functional layer of the model is the SincNet layer, followed by max pooling and layer normalization. The output of the SincNet layer for filter $i_{th}, i \in [1, N]$ in the SincNet layer is defined as follows:

$$\begin{aligned} h_i[n] &= x[n] * g[n, f_{i1}, f_{i2}] \\ &= x[n] * [2f_{i2}sinc(2\pi f_{i2}n) - 2f_{i1}sinc(2\pi f_{i1}n)] \end{aligned} \tag{1}$$

where $x[n]$ is the n_{th} chunk of the signal. $g[n, f_{i1}, f_{i2}]$ is used to represent the function of the i_{th} filter-bank. f_{i1} and f_{i2} are the low and high cut-off frequencies that need to be learned while training. The sinc function is defined as $sinc(x) = sin(x)/x$. To avoid the ripples in the passband and attenuation in the stop band, a Hamming window [26] is applied on $g[n, f_{i1}, f_{i2}]$. In Eq. 1, the filters are initialized with the cut off frequencies of the Mel-scale filter-bank, which has taken the human perception into consideration.

The second part of this layer is a standard 1-D convolutional layer, a max pooling layer and the layer normalization. The output $H[n]$ of the normalization layer is used as the input to the third part, the bidirectional LSTM, which can utilize both the forward and backward information of the input. Then, an attention layer and a dense layer are applied for feature weighting and mapping. The function of the attention layer is defined as:

$$\begin{aligned} u_t[n] &= tanh(Wh_t[n] + b) \\ \alpha_t &= \frac{exp(u_t[n]^T u)}{\sum_t exp(u_t[n]^T u)} \\ y[n] &= \sum_t \alpha_t h_t[n]. \end{aligned} \tag{2}$$

where $h_t[n]$ is the t_{th} component of $H[n]$ output by LSTM. $u_t[n]$ can be regarded as the hidden representation of $h_t[n]$ through a one-layer MLP. The importance of each hidden representation is measured by the normalized similarity between $u_t[n]$ 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" [27]. The system training is based on minimizing the loss between the predicted label and the ground-truth label. After training, the complete system can be regarded as the combination of the front-end feature extractor and the back-end feature classifier. To test the feature representation ability of

the feature extractor, as illustrated in Figure 1 we evaluated features extracted from either the dense layer or the attention layer. More details can be found in Section 4.3.

4. Experiment Setup

4.1. Dataset

The IVA dataset was collected at the Department of Neurology, University of Sheffield based at the Royal Hallamshire Hospital in a real clinical setting during 2016, 2017 and 2018 [10]. A *Digital Doctor* (or Intelligent Virtual Agent) presented on a laptop asks a series of conversational questions and administers a series of verbal tests. The questions are designed to mimic the neurologist-patient conversation happening as part of routine diagnostic assessments. The speech sampling rate is 16KHZ. In our experiment, only the audio recordings from the participants diagnosed with ND, MCI, and HCs are used. Further information about the data is given in Table 1 and in [10]. The average duration of the recordings in the IVA dataset is about 9 minutes which is too long to utilise directly as the input of the Sinc-CLA feature extractor. A similar problem is described in [8] and they chose to segment the input with manual information. As we are aiming for a fully automatic system, we instead chose to cut the recording into 2 second chunks. Each chunk is assigned a label corresponding to its diagnostic category.

Table 1: *IVA Dataset information; number of speakers, number of recordings and total duration for each diagnostic category.*

Diagnostic category	# Spk	# Rec	Duration
MCI	24	29	3h13min
ND	21	24	4h35min
HC	25	35	4h43min

4.2. Evaluation Setting

To provide a reliable result, 10 fold cross-validation (CV) is used on the relatively small dataset and each fold is fixed for all the experiments we present. The number of recordings in the three partitions (training, development, and test) of each fold is as balanced as possible in terms of the diagnostic category. In addition, as can be seen from Table 1, some speakers contributed more than one recording and these were kept in the same partition (speaker independent).

A typical classification pipeline is used to evaluate the extracted features. The front-end features are either the baseline feature sets or the features learned by Sinc-CLA, followed by the back-end classifier. Logistic Regression (LR) and Support Vector Machine (SVM), the most commonly used classifiers in acoustic-based cognitive decline detection fields, are adopted. The parameters in SVM are set as $C = 0.01$, kernel type=*rfb*. For each data fold, the features from training and development sets (9 folds) are used to train the back-end classifier and the test set is used for evaluation. The presented result is averaged across the 10 fold test set. Both the chunk-level and recording-level results are evaluated. The chunk-level result is calculated as majority voting over the predicted label output for each chunk by the classifier. To verify our system, the classification tasks include HCs vs. ND, HC vs. MCI and HC vs. people living with either ND or MCI.

4.3. Model Configuration

In the feature extraction part, the segmented chunks in the training set are fed into the designed feature extractor (Sinc-CLA). The SincNet layer is composed of N=80 filters of length L=125 samples . The parameters for the filters in the SincNet layer are initialized with the cut-off frequencies of the Mel-scale filterbank as introduced in [16]. The standard convolutional layer uses 60 filters of length 5. The max-pooling size of the two convolutional layers is 3. The number of units in the bidirectional LSTM is 50. The output of the Bidirectional LSTM layer is the 100 dimensional feature, which is the concatenation of the two 50 unidirectional LSTM outputs. The dimension of the attention matrix is set as 30. The output of the attention layer is the 100 dimension vector. The dense layer composes 1024 neuron units. In the model, all hidden layers use leaky-ReLU [28] non-linearities. $rmsprop$ [29] is applied as the optimizer with a learning rate of 0.01. While training, the mini-batch size is set to 30 and the epoch is set to 40. All the parameters of the network are selected according to the development set. F-measure is used as the criteria. After the feature extractor is trained, the 2 second chunks are input into the Sinc-CLA feature extractor. The features output by the attention layer and dense layer (named as 'attention feature' and 'dense feature' in the following) are used for the classification experiments.

4.4. Baseline Features

Research has shown promising results for using features initially proposed for emotion recognition in systems for automatic assessment of cognitive decline [8,30,31]. IS10 and ComParE features, which have achieved outstanding results [30,31], are adopted as the baseline feature sets in our experiment. The features are extracted by the OpenSMILE [25] toolkit. Compared with frame-level features, the statistic suprasegment feature can provide better performance on our task. To get the suprasegment feature for each 2 second chunk, the mean, maximum, minimum, median, and standard deviation are calculated across time on the frame-level feature matrix as in [9]. Then a list of 380 (76×5) features based on IS10 and 650 (130×5) features based on ComParE are generated.

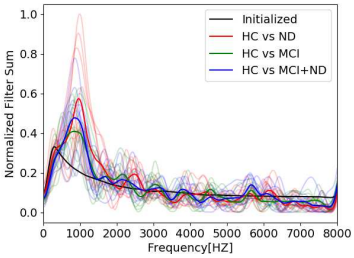
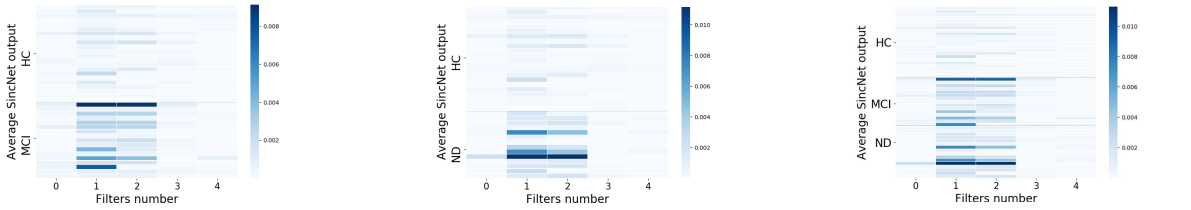


Figure 2: *Cumulative frequency response of SincNet filters on the three classification tasks. Bold lines are the average response for the 10-fold CV and thin lines are the response for every fold trained network.*

5. Results

5.1. Filter Analysis

Before describing the classification results, it is interesting to analyse the learned SincNet filters. Figure 2 shows the initialized and the three learned cumulative frequency responses



(a) The HC vs. ND classification task (b) The HC vs. MCI classification task (c) The HC vs. MCI+ND classification task

Figure 3: *The average representation of the SincNet output, only the recordings from the first fold training set are shown.*

(CFRs) of the SincNet layer. The black line corresponds to the initialised CFR, and the different coloured lines refer to different classification tasks after training.

The filter sum is normalized by the highest response. The conclusion can be summarized as:

1. Compared with the initialised CFRs, more details are shown in the learned CFRs. This shows that while training the filters, task specific information has been learned. It may also explain why Mel-scale filter bank based features are less suitable for our specific task.
2. By observing the CFRs of the three tasks, it can be seen that the frequency responses concentrate on the low frequencies, which is consistent with prior knowledge [6,9]. Though the low frequency information has been taken into consideration for some hand-crafted feature designation, they cannot achieve as good results as the features learned by our designed feature extractor (shown in Section 5.2).
3. Furthermore, compared with the other two tasks, the CFRs of the low frequency zone is higher for the HC vs. ND classifier. This may mean that for more severe symptoms, as seen in the ND cohort, more concentration should be put on low frequencies for classification.

The output of the SincNet layer is a $H \in [frame\_num \times filter\_num]$ matrix. As opposed to the CNN, the learned filters in a SincNet are ordered according to the frequency (from low to high, due to the Mel-scale initialization). The benefits of that is that the analysis of the SincNet output can help us better interpret the frequency related information which may be informative for cognitive decline assessment. To this end, the average $filter\_num$ (80) dimensional vector for each recording is calculated by averaging H over time. In Figure 3 only the first 5 out of 80 dimensions of the average vector is plotted as they are more distinctive. Each row corresponds to the filter response of one recording. As described in [32], the increase in the power of low frequency ranges can be a result of cognitive decline. The high values and main differences are concentrated in the first several filters for the three tasks.

5.2. Classification Result

The classification results on the baseline feature sets (IS10 and ComParE), and the dense and attention features are calculated by averaging across the 10 fold CV. Both the chunk-level and recording-level F-measure is calculated and presented in Table 2 and Table 3 respectively. In Table 2, comparing with the IS10 and ComParE feature sets, the classification results of the learned dense feature and attention feature are superior for the three classification tasks we performed. For example, for the HC vs. ND task, the best chunk-level classification performance is 88.39% achieved by dense feature classified by LR, compared with 81.34% achieved by IS10 classified by LR as the best baseline result. The performance of the dense and attention features do not differ much for either of the two classifiers. Comparing the two tables, the performance of the features and classifiers at the recording-level is better but consistent with the performance under the same situation after majority voting on the chunk-level labels.

Table 2: *The F-measure for chunk-level classification.*

Classifier	Feature	HC vs. ND	HC vs. MCI	MCI+ND vs. HC
LR	ComParE	77.08%	68.33%	75.15%
	IS10	81.34%	70.08%	77.51%
	dense	**88.39%**	78.19%	84.26%
	attention	88.15%	77.87%	84.18%
SVM	ComParE	72.58%	67.26%	70.70%
	IS10	78.28%	70.04%	75.64%
	dense	88.21%	**79.27%**	84.23%
	attention	88.35%	78.88%	**84.56%**

Table 3: *The F-measure for recording-level classification.*

Classifier	Feature	HC vs. ND	HC vs. MCI	MCI+ND vs. HC
LR	ComParE	88.09%	81.18%	81.60%
	IS10	93.25%	81.60%	84.31%
	dense	**98.29%**	84.09%	**93.18%**
	attention	96.58%	85.74%	**93.18%**
SVM	ComParE	89.83%	77.21%	77.13%
	IS10	93.25%	81.28%	82.57%
	dense	96.58%	85.61%	92.06%
	attention	96.58%	**87.27%**	**93.18%**

6. Conclusion

In this paper, a feature extractor named (Sinc-CLA) was designed for extracting task-driven features from the raw wave to classify recordings of people with neurodegenerative related disorders (ND, and HC). Compared with the IS10 and ComParE feature sets, the task-driven features achieved superior performance. Analyzing the CFRs of the SincNet layer gave us evidence that low-frequency information is critical for classifying MCI and ND from HC. The intuition of the learned filters and their output made the result more convincing.

7. Acknowledgement

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

[2] F. Jessen, B. Wiese, C. Bachmann, S. Eifflaender-Gorfer, F. Haller, H. Kölsch, T. Luck, E. Mösch, H. van den Bussche, M. Wagner *et al.*, "Prediction of dementia by subjective memory impairment: effects of severity and temporal association with cognitive impairment," *Archives of general psychiatry*, vol. 67, no. 4, pp. 414–422, 2010.

[3] H. Amieva, M. Le Goff, X. Millet, J. M. Orgogozo, K. Pérès, P. Barberger-Gateau, H. Jacqmin-Gadda, and J. F. Dartigues, "Prodromal Alzheimer's disease: successive emergence of the clinical symptoms," *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, vol. 64, no. 5, pp. 492–498, 2008.

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

[5] J. Chen, J. Zhu, and J. Ye, "An attention-based hybrid network for automatic detection of Alzheimer's disease from narrative speech," *Proc. Interspeech 2019*, pp. 4085–4089, 2019.

[6] J. J. G. Meilán, F. Martínez-Sánchez, J. Carro, D. E. López, L. Millian-Morell, and J. M. Arana, "Speech in Alzheimer's disease: Can temporal and acoustic parameters discriminate dementia?" *Dementia and Geriatric Cognitive Disorders*, vol. 37, no. 5-6, pp. 327–334, 2014.

[7] K. Lopez-de Ipiña, J. Alonso, J. Solé-Casals, N. Barroso, P. Henriquez, M. Faundez-Zanuy, C. Travieso, M. Ecay-Torres, P. Martinez-Lage, and H. Eguiraun, "On automatic diagnosis of alzheimer's disease based on spontaneous speech analysis and emotional temperature," *Cognitive Computation*, vol. 7, no. 1, pp. 44–55, 2015.

[8] 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.

[9] T. Alhanai, R. Au, and J. Glass, "Spoken language biomarkers for detecting cognitive impairment," in *Proc. of ASRU*, 2017.

[10] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.

[11] K. Horley, A. Reid, and D. Burnham, "Emotional prosody perception and production in dementia of the Alzheimer's type," *Journal of Speech, Language, and Hearing Research*, 2010.

[12] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[13] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform CLDNNs," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[14] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366, 1980.

[15] N. Zeghidour, N. Usunier, G. Synnaeve, R. Collobert, and E. Dupoux, "End-to-end speech recognition from the raw waveform," *arXiv preprint arXiv:1806.07098*, 2018.

[16] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 1021–1028.

[17] M. Sarma, P. Ghahremani, D. Povey, N. K. Goel, K. K. Sarma, and N. Dehak, "Emotion identification from raw speech signals using dnns." in *Interspeech*, 2018, pp. 3097–3101.

[18] Y. LeCun, P. Haffner, L. Bottou, and Y. Bengio, "Object recognition with gradient-based learning," in *Shape, contour and grouping in computer vision*. Springer, 1999, pp. 319–345.

[19] 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.

[20] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015.

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

[22] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.

[23] M. Ravanelli and Y. Bengio, "Interpretable convolutional filters with sincnet," *arXiv preprint arXiv:1811.09725*, 2018.

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

[25] 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.

[26] C. L. Clark, *LabVIEW digital signal processing*. Tata McGraw-Hill Education, 2005.

[27] 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.

[28] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. icml*, 2013.

[29] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural networks for machine learning*, vol. 4, no. 2, pp. 26–31, 2012.

[30] Y. Pan, B. Mirheidari, M. Reuber, A. Venneri, D. Blackburn, and H. Christensen, "Improving detection of Alzheimer's Disease using automatic speech recognition to identify high-quality segments for more robust feature extraction," *Interspeech*, 2020, submitted to Interspeech 2020.

[31] 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.

[32] 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.