Interspeech 2018
2-6 September 2018, Hyderabad

Automatic Early Detection of Amyotrophic Lateral Sclerosis from Intelligible Speech Using Convolutional Neural Networks

KwangHoon An[1], *Myungjong Kim*[1], *Kristin Teplansky*[1,2], *Jordan R. Green*[3], *Thomas F. Campbell*[2], *Yana Yunusova*[4], *Daragh Heitzman*[5], *Jun Wang*[1,2]

[1]Speech Disorders & Technology Lab, Department of Bioengineering
[2]Callier Center for Communication Disorders, University of Texas at Dallas, United States
[3]Department of Communication Sciences and Disorders
MGH Institute of Health Professions, United States
[4]Department of Speech-Language Pathology, University of Toronto, Canada
[5]MDA/ALS Center, Texas Neurology, United States

{kwanghoon.an, myungjong.kim, kristin.teplansky, wangjun}@utdallas.edu,
jgreen2@mghihp.edu, thomas.f.campbell@utdallas.edu, yana.yunusova@utoronto.ca,
dheitzman@texasneurology.com

Abstract

Amyotrophic lateral sclerosis (ALS) is a rapidly progressive neurodegenerative disease of the motor system that leads to the impairment of speech and swallowing functions. The lack of a biomarker typically causes a diagnostic delay. To advance the current diagnostic process, we explored the feasibility of automatic detection of patients with ALS at an early stage from highly intelligible speech. A speech dataset was collected from thirteen newly diagnosed patients with ALS and thirteen age- and gender-matched healthy controls. Convolutional Neural Networks (CNNs), including time-domain CNN and frequency-domain CNN, were used to classify the intelligible speech produced by patients with ALS and those by healthy individuals. Experimental results indicated both time- and frequency-CNN outperformed standard neural network. The best sample-level sensitivity and specificity were obtained by time-CNN (71.6% and 80.9%, respectively). When multiple samples were used to vote to estimate a person-level performance, the best result was obtained by frequency-CNN (76.9% sensitivity and 92.3% specificity). Results demonstrated the possibility of early detection of ALS from intelligible speech signals.

Index Terms: amyotrophic lateral sclerosis, human-computer interaction, computational paralinguistics

1. Introduction

Amyotrophic lateral sclerosis (ALS), also known as Lou Gehrig's disease, is a fatal and progressive motor neuron disease [1]. There is no definite diagnostic procedure and no cure for ALS. The current diagnosis of ALS is provisional, based primarily on clinical observations of upper and lower motor neuron damage in the absence of other causes [2]. Due to the lack of clinicopathologic markers of ALS, patients are often misdiagnosed (up to 45% of the time) or delayed for up to 12 months [3]. One unfortunate consequence of this delay is that by the time of diagnosis, a patient's motor neurons may have been affected already. ALS affects patient's bulbar system and then causes speech and swallowing problems [4]. The Amyotrophic Lateral Sclerosis Functional Rating Scale revised (ALSFRS-r), a self-report from patients/suspects, is currently used to predict ALS progression by analyzing the ability to complete functional activities in daily living [5]. The diagnosis and treatment of ALS will be significantly strengthened when objective, sensitive markers for the disease can be identified [6].

Recent studies indicated speech production decline is among the earliest indicators of bulbar motor involvement due to ALS [7, 8]. Dysarthria is a speech disorder resulting from deficits in musculature control of the articulators. Current common clinical measures for speech performance are subjective and non-deterministic at an early stage [8]. Perceptual analysis of dysarthric speech such as speech intelligibility (the percentage of words understood by listeners) and speaking rate (words per minute, WPM) are widely used methods to determine disease severity. However, symptoms of dysarthria may not be perceptually detectable until 80% of the motor neurons are loss [9]. Thus, perception-based approaches may not be able to distinguish speech produced by patients with early diagnosed ALS and healthy controls, because both of their speech are highly intelligible.

The automatic detection of other neurological diseases from speech signals recently has shown promising results for depression [10, 11, 12], traumatic brain injury [13], and Parkinson's disease [14, 15, 16, 17]. Various types of acoustic features, such as formant centralization ratio, vowel space area, intonation, and prosody [18, 19] have been used for the detection of neurological diseases. One advantage of using speech signals is that speech samples can be easily obtained from subjects (e.g., using smartphone application) without the logistical difficulty in a clinical environment.

Our previous preliminary study demonstrated that speech may be a sensitive measure to automatically detect ALS at early stage and monitor disease progression [6, 20]. Our previous work on early detection of ALS [6], however, used a dataset that contains the non-age-matched healthy controls, which may introduce a bias in the classification performance. In comparison to younger speakers, elderly individuals may have slower and breathy voice characteristics. In addition, both speakers with ALS and senior-aged individuals show slow speech and breathy voice. Thus, further work with an age- and gender-matched dataset is needed to verify the previous findings.

In this study, we improved our previous design [6] by using a larger dataset with age- and gender-matched healthy controls. Similar to our previous work, the current work includes data from early-diagnosed patients with ALS who still produce

10.21437/Interspeech.2018-2496

Table 1: *Patients information statistics*

Subject	Age	Speaking Rate	Speech Intell.
A01	55	235.7	100
A02	52	164.2	99
A03	61	209.5	96.4
A04	54	192.4	99.1
A05	42	167.5	97.3
A06	58	180.8	100
A07	60	167.6	95.5
A08	56	189.2	100
A09	42	222.2	98.2
A10	54	156	100
A11	48	155.6	99
A12	61	161.4	97.3
A13	48	217.8	100
Mean	53.9	186.1	98.6
Std	6.4	27.4	1.6

Table 2: *Healthy controls information statistics*

Statistics	Age	Speaking Rate	Speech Intell.
Mean	63.5	189.8	99.86
Std	8.7	16.5	0.3

highly intelligible speech, although some of them have shown limb symptoms.

In addition to predefined, hand-crafted features and artificial neural network (ANN) that were used in our previous work [6], we applied convolutional neural network (CNN)-based representation learning in the current work. Representation learning can learn useful features from low-level signals and it has shown effectiveness in various classification applications, outperforming traditional hand-crafted features [21, 22]. In particular, CNNs are one of the widely used representation learning methods due to the ability of extracting local features through convolution and pooling operations [23]. With time-frequency signals, there are various types of CNNs such as time-domain CNN and frequency-domain CNN depending on the convolution and pooling operations along the time and frequency axes, respectively. CNNs can extract robust features across temporal and spectral domains, which has shown its effects in other audio signals based studies [24, 25, 26]. Thus, the use of CNN-based representation learning may have a benefit in extracting useful information for ALS detection. In this paper, we tested CNN-based representation learning approaches on low-level filterbank energies with time-domain and frequency-domain CNNs to detect early-stage ALS disease using intelligible speech samples. The performance of CNNs were compared with hand-crafted features and ANN, the previous approach [6]. A leave-one-paired-subject-out cross validation strategy was used to evaluate the performance of these classification approaches.

2. Dataset

The speech data set used in this study was collected from thirteen early-diagnosed patients with ALS (9 females and 4 males) and thirteen healthy age-matched speakers (8 females and 5 males). The patients with ALS were diagnosed within 6-12 months prior to data collection. The age interval of the patients is from 42 to 61 (mean = 53.9, SD = 6.4). The age range of the healthy controls is from 47 to 73 (mean = 63.5, SD = 8.6). The patient and healthy control information is summarized in Table 1 and 2. As previously mentioned, speech intelligibility is a perceptual measurement of speech clarity. Speaking rate is the amount of words produced per minute. A certified speech-language pathologist evaluated the speech intelligibility and speaking rate of the participants using the Sentence Intelligibility Test (SIT) software [27]. As shown in Table 1, both individuals with ALS and healthy controls had high speech intelligibility levels and similar speaking rates.

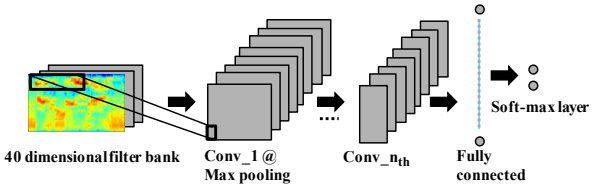

Figure 1: *Example of the time-CNN architecture.*

At each data collection session, acoustic data were collected during the production of 20 sentences such as *I need some assistance* and *call me back when you can*. Each sentence was produced 4 times for a total of 80 productions. The stimuli were selected because they are often used in alternative and augmentative communication (AAC) devices [6]. The built-in microphone in the NDI Wave system [28] was used to collect acoustic data. Articulatory movement data (i.e., tongue and lips) were also collected for future studies. All the speech data had a 16 kHz sampling rate. A total of 2,080 valid speech samples were collected, where each sample is an acoustic sample of a spoken short phrase.

3. Method

We performed ALS classification from speech samples using the following three classification methods: 1) ANN with statistical (hand-crafted) features as in our previous work [6], 2) time-domain CNN, and 3) frequency-domain CNN based representation learning approaches.

3.1. Baseline Approach: ANN with Statistical Features

Our previous work [6] used statistical features extracted by the publicly available tool, openSMILE. We set up our previous work as a baseline and call this approach as ANN with statistical features. In the baseline approach, openSMILE [29] was used to extract hand-crafted features that are statistical variation of the widely used acoustic features, such as the mean and standard deviation of mel-frequency cepstral coefficients (MFCCs) and the quartile of the fundamental frequency contour, from the speech samples. Total 7,755 acoustic features were extracted from each speech sample and fed into the ANN-based classification model. The ANN has the 2 dimensional softmax output layer: *ALS* and *healthy*. The test samples were classified by a maximum a posteriori probability obtained from ANN. Although the network implementation was based on the ANN framework in TensorFlow [30], it had only one hidden layer. Thus, we still call the model ANN throughout this paper.

3.2. CNN with Filterbank Energies

Representation learning is a feature learning method in which the model learns useful feature representation from low-level signals without hand-crafted feature extraction. For example, MFCC, one of the widely used hand-crafted features in speech recognition, includes de-correlation of filterbank energies in spectral domain. The de-correlation may lead to loss of use-

ful information to discriminate ALS and healthy from speech signals.

Convolutional neural networks (CNNs) consist of convolution and pooling layers. The convolution operation is the dot product between the small region of the input and the localized weights. The localized weights are shared by convolving entire input. As an output of the convolution, learned features are obtained in the form of the multi-dimensional feature maps. Each feature map represents localized information of the input. Then, max-pooling operation is applied to the features to reduce the dimension by selecting the maximum value out of all the features of the small region. Thus, CNN has benefit of extracting useful small variation-insensitive localized features from low-level signals.

In this work, we used CNN-based representation learning approaches. We used 40 low-level filterbank energies with 16 ms frame and its first derivative (delta) and second derivative (delta-delta) as an input to the model. The filterbank, delta, and delta-delta are concatenated to form 3D channel shape and fed into the model. Figure 1 illustrates an example of the CNN architecture with n layers. Depending on the direction of convolution, we considered two types of CNNs, time-domain CNN and frequency-domain CNN.

3.2.1. Time-CNN

Time-domain convolution applies convolution and pooling operations over time, and therefore, it can extract modulating characteristics while keeping invariance to a small shift in time [24]. Three layers of CNNs were used with different filter sizes of 1×6, 1×5, and 1×3 for the corresponding layers, respectively. Each layer was sub-sampled by non-overlapping max pooling operation with 1×2, 1×3, and 1×3 with 64 feature maps, respectively.

3.2.2. Frequency-CNN

Frequency-domain convolution applies convolution and pooling operations along frequency, and therefore, it can represent useful spectral features while reducing frequency variance [24]. Three layers of CNNs were used with different filter sizes of 7×1, 5×1, and 3×1 for the corresponding layers, respectively. Each layer was sub-sampled by non-overlapping max pooling operation with 2×1, 4×1, and 4×1 with 64 feature maps, respectively.

3.3. Experimental Design

The ANN to train statistical features had 1 hidden layers with 3072 hidden neurons and binary output softmax layer. We tested from 1 to 4 layers with 256, 512, 1024, 2048, 3072, 4096 neurons at each layer and obtained the best result with 1 hidden layer with 3072 hidden neurons. Each hidden neuron was activated by the rectified linear unit (ReLU). The Adam optimizer [31] was employed for training with backpropagation.

For each type of CNNs, we tested 4 to 8 filter sizes in the first layer and reduced size by one to three in the following layers. Feature representation at the last CNN layer is flattened and fed into a fully-connected layer with 256 hidden neurons .

As mentioned previously, a total of 2,080 acoustic samples were collected during the data collection, where each participant produced 80 acoustic samples. We paired 80 samples from one patient and 80 samples from one healthy control as a test set and performed leave-one-subject-pair-out cross validation to evaluate the performance of the ANN + statistical features and CNN-based representation learning method in a speaker-independent way. In the CNN-based representation learning, we used fragmented context windows as single input within an utterance, which will produce multiple maximum a posteriori at the output layer of the model. In the inference step of the CNN based method, we determined predicted label of single utterance sample by averaging the obtained probabilities of all context windows within one utterance.

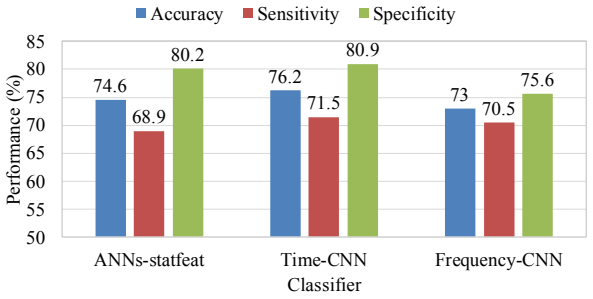

Figure 2: *Performance of ALS detection of the three approaches: ANN with statistical features, time-CNN, and frequency-CNN.*

Accuracy, sensitivity, and specificity were the major performance indicators in this experiment. Accuracy is the overall probability of correctly classified samples over the total number of samples. Sensitivity is the probability of correctly predicted acoustic samples as a patient given all patient samples. Specificity is the probability of correctly classified healthy controls samples given all healthy control samples. Accuracy, sensitivity, and specificity are calculated as follow:

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)} \quad (1)$$

$$Sensitivity = \frac{TP}{(TP + FN)} \quad (2)$$

$$Specificity = \frac{TN}{(TN + FP)} \quad (3)$$

where TP is true positive, TN is true negative, FP is false positive, and FN is false negative. Here, positive means a prediction that the speech sample is produced by a speaker with ALS; negative means a prediction that the speech sample is produced by a healthy control.

4. Result and Discussion

Figure 2 shows the accuracy, sensitivity, and specificity of the ALS classification using the three approaches: ANN with statistical features, time-CNN and frequency-CNN. Also, a two-tailed t-test was performed to measure if there were statistical significance between the performance of the three approaches and random guess. ANN with statistical features, time-CNN, and frequency-CNN achieved accuracies of 74.6% ($p < 0.01$), 76.2% ($p < 0.001$), and 73.0% ($p < 0.001$), respectively. The three approaches achieved sensitivity of 68.9% ($p = 0.07$), 71.5% ($p < 0.05$), and 70.5% ($p < 0.05$), respectively. The three approaches achieved specificity of 80.2% ($p < 0.001$), 80.9% ($p < 0.00001$), and 75.6% ($p < 0.0001$), respectively.

All of the three measures by the three approaches were significantly above chance level (50%) ($p < 0.05$), except the sensitivity predicted by ANN. All the measures predicted using time-CNN and frequency-CNN were above chance level. These

Table 3: *Individual subject's performance in cross validations (CV).*

	ANN+statfeat		Freq-CNN		Time-CNN	
CV	Sens	Spec	Sens	Spec	Sens	Spec
A01-H01	100	97.5	96.3	82.5	100	95
A02-H02	100	66.3	82.5	78.8	82.5	77.5
A03-H03	100	73.8	86.3	100	96.3	96.3
A04-H04	42.5	77.5	23.8	68.8	22.5	57.5
A05-H05	92.5	100	86.3	81.3	95	97.5
A06-H06	100	98.8	88.8	77.5	98.8	78.8
A07-H07	100	87.5	100	72.5	100	72.5
A08-H08	23.8	92.5	63.8	87.5	77.5	91.3
A09-H09	30	93.8	86.3	76.3	75	70
A10-H10	8.7	48.8	25	38.8	20	81.3
A11-H11	58.7	18.8	57.5	62.5	41.3	67.5
A12-H12	48.8	96.2	45	85	47.5	100
A13-H13	91.3	91.3	75	71.3	73.8	66.3
Mean	68.9	80.2	70.5	75.6	71.6	80.9
Std	34.5	23.8	25.7	14.5	29.2	13.9

results showed the feasibility to automatically detect ALS from healthy controls using speech signals, which further verified the finding of our previous work [6].

In addition, our result showed that both time-CNN and frequency-CNN outperformed both ANN with statistical features. The sensitivity of the time-CNN is 71.6%, whereas the sensitivity of the ANN with statistical features and the frequency-CNN are 68.9% and 70.5%, respectively. The specificity of time-CNN, frequency-CNN, and ANN are 80.9%, 75.6%, and 80.2%, respectively.

Table 3 gives the detailed classification results for each subject-pair in each cross validation using the three approaches. The mean and standard deviation of the performance of the validations were given in the lower part of the table. Despite the promising results obtained, a large standard deviation across each subject-pair may suggest a person-level classification that is lower than expected.

In practice, diagnosis for each ALS suspect/candidate can be determined by a vote of the detection results of multiple speech samples. For example, each ALS suspect can be asked to produce five different speech (phrase) samples. Each sample will be fed into one classifier that makes a prediction if a suspect is with ALS or not. If more than half (three or more) samples indicate the suspect is with ALS, then the suspect is predicted as ALS; otherwise, the suspect is classified as healthy.

To estimate the person-level classification, we used the probability (p in Eq. 4) of detecting ALS from each acoustic sample that follows our best sample-level performance by time-CNN (with 76.2% accuracy, 71.5% sensitivity, and 80.9% specificity). The suspect will produce total N different acoustic samples and the diagnosis will be determined by the probability of k correctly classified samples or greater than integer k, where $k = \lfloor \frac{N}{2} \rfloor + 1$. The probability can be calculated by cumulative binomial distribution [32] as:

$$P(X \geq k) = 1 - \sum_{i=0}^{k-1} \binom{n}{i} \cdot p^i (1-p)^{n-i} \qquad (4)$$

Figure 3 shows the results of the estimated person-level classification estimated using Eq. 4. and the actual test results based on sample voting. With the assumption of binomial distribution of the sample-level classification, when a suspect provides five samples, the accuracy, sensitivity, and specificity could be up to 90.8%, 85.6%, and 94.9%, respectively. When a suspect provides 15 samples, the accuracy, sensitivity, and specificity could be up to 98.7%, 96.3%, and 99.7%, respectively. However, the best actual person-level performances are 84.6% accuracy, 76.9% sensitivity, and 92.3% sensitivity, when we took all 80 samples for a voting using Frequency-CNN. Time-CNN obtained a 80.8% accuracy, 69.3% sensitivity, and 92.3% specificity. The finding indicated that sample-level performance may not match a binomial distribution (possibly due to the small number of subjects).

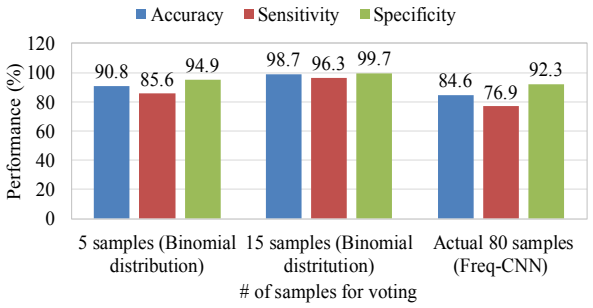

Figure 3: *Person-level ALS detection performance.*

Limitation. Despite the promising results, a large variation of person-level detection was also observed. A dataset from a larger number of subjects (both ALS and healthy subjects) is needed to verify these findings. In addition, in this work, we employed the data only from ALS patients and healthy controls, without including data from other diseases. Because other diseases also cause dysarthric speech, there could be a possibility that our model has been trained to discriminate between potential dysarthria from non-dysarthria (although both are perceptionally intelligible speech). Verification of this needs a dataset with speech samples from patients with other diseases (e.g., Parkinson's disease[16]) and age-matched healthy controls.

5. Conclusions and Future Work

In this project, we explored the feasibility of automatic early detection of ALS from highly intelligible speech from age-matched healthy controls using CNNs. Results showed that the CNNs outperformed traditional approaches (e.g., standard neural network) with statistical features measured by sensitivity and specificity. Time-CNN outperformed frequency-CNN at sample-level prediction, while frequency-CNN slightly outperformed time-CNN at person-level prediction.

Future work will include other dysarthric-related diseases to exclude a potential bias caused by dysarthria and focus on detection of ALS disease in the absence of other factors. Another future direction is to add articulatory information on top of the acoustic information to advance the detection. Other deep learning models (e.g., recurrent neural networks) will be also investigated.

6. Acknowledgements

This work was supported by the National Institutes of Health through grants R03DC013990 and R01DC013547, and the American Speech-Language-Hearing Foundation through a New Century Scholar grant. We would like to thank Jennifer McGlothlin, Alyssa Shrode, Bjorn Bleta, Brittany Shrode, and the volunteering participants.

7. References

[1] M. C. Kiernan, S. Vucic, B. C. Cheah, M. R. Turner, A. Eisen, O. Hardiman, J. R. Burrell, and M. C. Zoing, "Amyotrophic lateral sclerosis," *The Lancet*, vol. 377, no. 9769, pp. 942–955, 2011.

[2] B. R. Brooks, R. G. Miller, M. Swash, and T. L. Munsat, "El Escorial revisited: Revised criteria for the diagnosis of amyotrophic lateral sclerosis," *Amyotrophic Lateral Sclerosis and Other Motor Neuron Disorders*, vol. 1, pp. 293–299, 2000.

[3] Y. Iwasaki, K. Ikeda, and M. Kinoshita, "The diagnostic pathway in amyotrophic lateral sclerosis," *Amyotrophic Lateral Sclerosis and Other Motor Neuron Disorders*, vol. 2, no. 3, pp. 123–126, 2001.

[4] S. E. Langmore and M. Lehman, "The orofacial deficit and dysarthria in ALS," *Journal of Speech and Hearing Research*, vol. 37, pp. 28–37, 1994.

[5] J. M. Cedarbaum, N. Stambler, E. Malta, C. Fuller, D. Hilt, B. Thurmond, A. Nakanishi, B. A. S. Group, A. complete listing of the BDNF Study Group *et al.*, "The alsfrs-r: a revised als functional rating scale that incorporates assessments of respiratory function," *Journal of the neurological sciences*, vol. 169, no. 1-2, pp. 13–21, 1999.

[6] J. Wang, P. V. Kothalkar, B. Cao, and D. Heitzman, "Towards automatic detection of amyotrophic lateral sclerosis from speech acoustic and articulatory samples," in *INTERSPEECH*, 2016, pp. 1195–1199.

[7] J. R. Green, Y. Yunusova, M. S. Kuruvilla, J. Wang, G. L. Pattee, L. Synhorst, L. Zinman, and J. D. Berry, "Bulbar and speech motor assessment in ALS: Challenges and future directions," *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, vol. 14, pp. 494–500, 2013.

[8] K. M. Allison, Y. Yunusova, T. F. Campbell, J. Wang, J. D. Berry, and J. R. Green, "The diagnostic utility of patient-report and speech-language pathologists ratings for detecting the early onset of bulbar symptoms due to ALS," *Amyotrophic Lateral Sclerosis and and Frontotemporal Degeneration*, vol. 18, pp. 358–366, 2017.

[9] B. Tomik and R. J. Guiloff, "Dysarthria in amyotrophic lateral sclerosis: a review," *Amyotrophic Lateral Sclerosis*, vol. 11, no. 1-2, pp. 4–15, 2010.

[10] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015.

[11] T. F. Quatieri and N. Malyska, "Vocal-source biomarkers for depression: A link to psychomotor activity," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[12] Z. Liu, B. Hu, L. Yan, T. Wang, F. Liu, X. Li, and H. Kang, "Detection of depression in speech," in *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE, 2015, pp. 743–747.

[13] M. Falcone, N. Yadav, C. Poellabauer, and P. Flynn, "Using isolated vowel sounds for classification of mild traumatic brain injury," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7577–7581.

[14] J. C. Vásquez Correa, J. R. Orozco Arroyave, J. D. Arias-Londoño, J. F. Vargas Bonilla, and E. Noth, "New computer aided device for real time analysis of speech of people with parkinson's disease," *Revista Facultad de Ingeniería Universidad de Antioquia*, no. 72, pp. 87–103, 2014.

[15] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of parkinson's disease," *IEEE Transactions on biomedical engineering*, vol. 59, no. 5, pp. 1264–1271, 2012.

[16] S. Hahm and J. Wang, "Parkinson's condition estimation using speech acoustic and inversely mapped articulatory data," in *INTERSPEECH*, 2015, pp. 513–517.

[17] E. Vaiciukynas, A. Verikas, A. Gelzinis, and M. Bacauskiene, "Detecting parkinsons disease from sustained phonation and speech signals," *PloS one*, vol. 12, no. 10, p. e0185613, 2017.

[18] S. Sapir, L. O. Ramig, J. L. Spielman, and C. Fox, "Formant centralization ratio: A proposal for a new acoustic measure of dysarthric speech," *Journal of speech, language, and hearing research*, vol. 53, no. 1, pp. 114–125, 2010.

[19] S. Skodda, W. Grönheit, and U. Schlegel, "Intonation and speech rate in parkinson's disease: General and dynamic aspects and responsiveness to levodopa admission," *Journal of Voice*, vol. 25, no. 4, pp. e199–e205, 2011.

[20] J. Wang, P. V. Kothalkar, M. Kim, Y. Yunusova, T. F. Campbell, D. Heitzman, and J. R. Green, "Predicting intelligible speaking rate in individuals with amyotrophic lateral sclerosis from a small number of speech acoustic and articulatory samples," in *Workshop on Speech and Language Processing for Assistive Technologies*, vol. 2016. NIH Public Access, 2016, pp. 91–97.

[21] G. Antipov, S.-A. Berrani, N. Ruchaud, and J.-L. Dugelay, "Learned vs. hand-crafted features for pedestrian gender recognition," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 1263–1266.

[22] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[23] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[24] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 22, no. 10, pp. 1533–1545, 2014.

[25] T. N. Sainath, B. Kingsbury, A.-r. Mohamed, G. E. Dahl, G. Saon, H. Soltau, T. Beran, A. Y. Aravkin, and B. Ramabhadran, "Improvements to deep convolutional neural networks for lvcsr," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 315–320.

[26] L. Tóth, "Combining time-and frequency-domain convolution in convolutional neural network-based phone recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 190–194.

[27] D. R. Beukelman, K. M. Yorkston, M. Hakel, and M. Dorsey, "Speech Intelligibility Test (SIT) [Computer Software]," 2007.

[28] J. R. Green, J. Wang, and D. L. Wilson, "Smash: a tool for articulatory data processing and analysis." in *Interspeech*, 2013, pp. 1331–1335.

[29] 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*. ACM, 2013, pp. 835–838.

[30] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, "Tensorflow: A system for large-scale machine learning," in *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, 2016, pp. 265–283.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, p. 15 pages, 2015.

[32] G. P. Wadsworth, *Introduction to Probability and Random Variables*. New York: McGraw-Hill, 1960.