AUTOMATIC DYSARTHRIC SPEECH DETECTION EXPLOITING PAIRWISE DISTANCE-BASED CONVOLUTIONAL NEURAL NETWORKS

Parvaneh Janbakhshi[1,2], *Ina Kodrasi*[1], *Hervé Bourlard*[1,2]

[1]Idiap Research Institute, Speech and Audio Processing Group, Martigny, Switzerland
[2]École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
{parvaneh.janbakhshi,ina.kodrasi,herve.bourlard}@idiap.ch

ABSTRACT

Automatic dysarthric speech detection can provide reliable and cost-effective computer-aided tools to assist the clinical diagnosis and management of dysarthria. In this paper we propose a novel automatic dysarthric speech detection approach based on analyses of pairwise distance matrices using convolutional neural networks (CNNs). We represent utterances through articulatory posteriors and consider pairs of phonetically-balanced representations, with one representation from a healthy speaker (i.e., the reference representation) and the other representation from the test speaker (i.e., test representation). Given such pairs of reference and test representations, features are first extracted using a feature extraction front-end, a frame-level distance matrix is computed, and the obtained distance matrix is considered as an image by a CNN-based binary classifier. The feature extraction, distance matrix computation, and CNN-based classifier are jointly optimized in an end-to-end framework. Experimental results on two databases of healthy and dysarthric speakers for different languages and pathologies show that the proposed approach yields a high dysarthric speech detection performance, outperforming other CNN-based baseline approaches.

Index Terms— Dysarthria, Parkinson's disease, Amyotrophic Lateral Sclerosis, pairwise distance, convolutional neural network

1. INTRODUCTION

Dysarthria is a commonly occurring speech disorder arising from brain damage associated with several neurological diseases such as Parkinson's disease (PD) and Amyotrophic Lateral Sclerosis (ALS). Since several components of the speech production mechanism are disrupted, dysarthria results in impaired phonation, prosody, and articulation [1]. For diagnosis, management, and treatment of patients, these impaired speech dimensions need to be evaluated through clinical auditory-perceptual assessments. However, such clinical assessments are subjective, time-consuming, and inefficient [2, 3].

To assist the clinical diagnosis of dysarthria and to avoid the drawbacks associated with clinical assessments, automatic dysarthric speech detection techniques can be used [3]. Typical automatic techniques developed in the research community are based on i) handcrafting acoustic features characterizing different impaired speech dimensions and ii) training classifiers using the handcrafted acoustic features to discriminate between dysarthric and healthy speech [4, 5]. Many acoustic features have been exploited to characterize impaired speech dimensions among which are Mel frequency cepstral coefficients, spectro-temporal sparsity parameters and rhythm-based metrics [6–13]). Although typical contributions for automatic dysarthric speech detection are based on such features, handcrafted acoustic features may fail to characterize abstract (but similarly important) acoustic cues that can further assist in differentiating dysarthric speech from healthy speech.

Seeking to exploit high-level abstract representations, there has been a growing interest in the research community to leverage data-driven deep learning approaches [14–18]. While deep learning approaches have dramatically improved the state-of-the-art in many speech processing applications, their advantages are yet to be established in the field of pathological speech assessment [19]. The main challenge in successfully exploiting deep learning approaches in pathological speech assessment is alleviating overfitting issues associated with the typically limited training data that is available.

To increase the number of training samples in [14–16], speech signals are split into short segments (e.g., 160 ms), each segment is labeled as healthy or dysarthric depending on the label of the complete signal, and convolutional neural networks (CNNs) are trained on these segments for dysarthric speech detection. Although considering short segments increases the number of training samples available per speaker, such short segments do not always exhibit dysarthric characteristics, and the CNNs are not guided to ignore speaker variabilities that are unrelated to dysarthria. A similar approach is also used in [18] where cascaded CNN and long short-term memory (LSTM) layers are exploited to classify the speech segments. In [17], LSTM Siamese networks are used for dysarthric speech detection. Networks with Siamese architectures are trained on pairs of input data with the same phonetic content. Pairwise training is advantageous when limited training data is available, since it guides the network to extract features that are discriminative of dysarthria while being robust to other unrelated speaker variabilities. However, since input data needs to have the same phonetic content, different LSTM networks need to be trained for different utterances.

Instead of using short speech segments to augment training data, we propose to use a CNN-based dysarthric speech detection system exploiting pairwise distance matrices. While our system benefits from pairwise training, a single network can be used for different utterances since the CNN operates on distance matrices instead of pairs of input data as in [17]. Inspired by the CNN-based query detection system in [20], we consider utterances from healthy speakers as reference representations, and we propose to compute frame-level distance matrices between these reference representations and phonetically-balanced test representations. We hypothesize that when the test speaker is healthy, the pattern of the distance

The authors would like to acknowledge the support of the Swiss National Science Foundation project no CRSII5_173711 "MoSpeeDi" on "*Motor Speech Disorders: characterizing phonetic speech planning and motor speech programming/execution and their impairments*".

matrix between the test and reference (i.e., healthy) representations is different (i.e., it is expected to be more quasi-diagonal) than when the test speaker is dysarthric. This distance pattern can be used as the input to a CNN-based binary classifier, which then categorizes it as an example from a healthy speaker (i.e., the distance pattern arises from comparing a healthy utterance to the reference representation) or as an example from a dysarthric speaker (i.e., the distance pattern arises from comparing a dysarthric utterance to the reference representation). Although such a CNN can directly operate on distance matrices computed from user-defined representations of utterances (e.g., the short-time Fourier transform (STFT)), these user-defined representations might not be optimal for healthy and dysarthric speech detection. To ensure that distance matrices are computed on optimal representations for our task, we propose to incorporate a front-end feature extraction layer to the network prior to computing distance matrices. The front-end feature extraction layer, the distance matrix computation, and the final healthy and dysarthric speech detection layers are jointly optimized in an end-to-end learning framework.

For the user-defined utterance representations, we propose to use articulatory posteriors (APs) instead of the STFT representation used in [14]. The use of APs is motivated by their potential to characterize articulation deficits in dysarthria, their robustness to noise, and their multilingual and cross-lingual portability [21]. Experimental results on two databases of Spanish and French healthy and dysarthric speakers show the advantages of using AP representations in comparison to STFT representations. Further, experimental results show that the proposed pairwise distance-based CNN with front-end feature extraction can yield a high dysarthric speech detection accuracy, also outperforming a baseline CNN system adapted from [14] and a pairwise distance-based CNN without front-end feature extraction.

2. TECHNICAL APPROACH

Fig.1 depicts a schematic representation of the proposed pairwise distance-based dysarthric speech detection CNN. As shown in this figure, the input to the system consists of pairs of reference and test representations of utterances. We follow the same procedure as in [22] to extract AP features for the representations of utterances (cf. Section 3.2). These representations are transformed through a feature extraction block prior to computing the distance matrix. The distance matrix is then considered as an image by a CNN-based classifier as in a standard binary image classification task. The complete architecture is optimized in an end-to-end framework to achieve dysarthric speech detection.

In the following, we present details on the different components of the proposed system, i.e., i) the front-end feature extraction, ii) the distance matrix computation, and iii) the CNN-based classifier.

2.1. Front-end feature extraction

We consider pairs of phonetically-balanced AP representations of utterances from two speakers; one utterance being a reference representation from a healthy speaker and the other utterance being from a test (healthy or dysarthric) speaker. Let us denote by $\mathbf{R}$ the $(F_1 \times M)$–dimensional reference representation, with F_1 being the number of AP features and M being the number of time frames in the reference representation. Similarly, let us denote by $\mathbf{T}$ the $(F_1 \times N)$–dimensional test representation, with N being the number of time frames in the test representation. To be able to handle variable-length inputs, we fix the length of all representations to a predetermined (user-defined) size S as in [20]. Representations with more time frames than S, i.e., $M > S$ or $N > S$, are down-sampled by deleting time frames in regular intervals. Representations with less time frames than S, i.e., $M < S$ or $N < S$, are padded at the beginning and end with time frames filled with a constant value. The constant value is arbitrarily set to the maximum value in the representation. We denote the resized reference and test representations by $\mathbf{R}_s$ and $\mathbf{T}_s$ and hypothesize that they contain similar (healthy or dysarthria-related) cues as in the original representations $\mathbf{R}$ and $\mathbf{T}$.

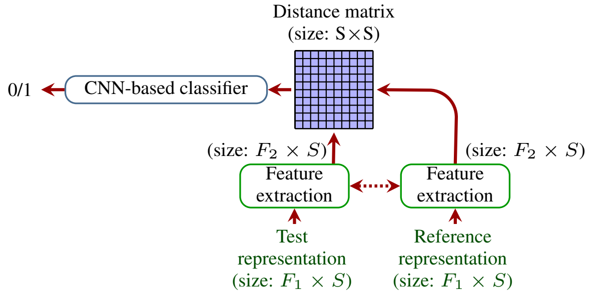

Fig. 1: Block diagram of the proposed pairwise distance-based dysarthric speech detection CNN. The two feature extraction blocks share the same set of parameters.

The front-end feature extraction block transforms the $(F_1 \times S)$–dimensional representations $\mathbf{R}_s$ and $\mathbf{T}_s$ into $(F_2 \times S)$–dimensional representations. To this end, we use a 1D convolution layer with F_2 channels such that the F_1–dimensional AP feature vectors for each time frame are transformed into F_2–dimensional feature vectors. Since this layer is jointly optimized with the distance matrix computation (cf. Section 2.2) and the CNN-based classifier (cf. Section 2.3) in an end-to-end framework, it can be expected that the transformed $(F_2 \times S)$–dimensional representations are more discriminative representations for the dysarthric speech detection task.

The architecture of the front-end layer is summarized in Table 1, where we have used $F_2 = 32$. It should be noted that the parameters of the front-end feature extraction layer to compute both test and reference feature representations are the same (cf. Fig. 1).

2.2. Distance matrix computation

The distance matrix is computed from the representations at the output of the feature extraction block. Let us denote the reference representation after feature extraction by $\hat{\mathbf{R}} = [\mathbf{r}_1, \ldots, \mathbf{r}_S]$, with $\mathbf{r}_i$, $i = 1, \ldots, S$, being the F_2–dimensional feature vector at time frame i. Similarly, the test representation after feature extraction is denoted by $\hat{\mathbf{T}} = [\mathbf{t}_1, \ldots, \mathbf{t}_S]$, with $\mathbf{t}_j$, $j = 1, \ldots, S$, being the F_2-dimensional feature vector at time frame j. The frame-level distance matrix $\mathbf{D}$ between the representations $\hat{\mathbf{T}}$ and $\hat{\mathbf{R}}$ is an $(S \times S)$–dimensional matrix, where the (i, j)–th entry is computed as the distance d between $\mathbf{t}_i$ and $\mathbf{r}_j$, i.e.,

$$\mathbf{D}_{i,j} = d(\mathbf{t}_i, \mathbf{r}_j). \tag{1}$$

Table 1: *Front-end feature extraction architecture.*

Layer	Description
Input	Size: (1xF_1xS): input speech representation
Conv1d + Relu	Channel: in=1, out=32, Filter: F_1x1, Stride: 1

To compute $\mathbf{D}$ within the proposed end-to-end framework, Euclidean distance is used, i.e., $d(\mathbf{t}_i, \mathbf{r}_j) = ||\mathbf{t}_i - \mathbf{r}_j||$. Since the reference representation $\hat{\mathbf{R}}$ always belongs to a healthy speaker, we expect the pattern of the so-computed distance matrix $\mathbf{D}$ to be more quasi-diagonal (i.e., contain more zeros on the diagonal due to similar $\mathbf{t}_i$ and $\mathbf{r}_j$) when the test representation $\hat{\mathbf{T}}$ belongs to a healthy speaker than when it belongs to a dysarthric speaker.

2.3. CNN-based classifier with pairwise distance matrices

The distance matrices computed in Section 2.2 serve as input to our CNN classifier. As summarized in Table 2, the CNN classifier consists of two 2D convolutional layers, followed by two Maxpooling and two fully connected (FC) layers. To prevent overfitting, dropout is employed during training. The label for each distance matrix fed into the CNN classifier is the label of the test speaker (healthy or dysarthric) used for the distance matrix computation.

The classifier is trained using distance matrices computed from all phonetically-matched pairs of test and reference representations in the training set. As mentioned in Section 1, a single network can be used for different utterances since the CNN operates on distance matrices instead of pairs of input data as in [17]. To evaluate an utterance from an unseen test speaker, we pair it to its phonetically-matched counterpart from many reference speakers in the training set and compute multiple distance matrices. All available distance matrices are then independently processed by the CNN classifier, and the final decision for the unseen test speaker is made by applying soft voting on all CNN prediction scores for all available distance matrices from that speaker.

3. EXPERIMENTAL RESULTS

In this section, the performance of the proposed pairwise distance-based dysarthric speech detection system is evaluated and compared to baseline systems.

3.1. Databases

To investigate the applicability and generalisability of the proposed approach to different pathologies and languages, two databases are considered.

PC-GITA database [23]. We consider Spanish recordings from 50 PD patients (25 males, 25 females) and 50 healthy speakers (25 males, 25 females) from the PC-GITA database [23]. Each speaker utters 24 words, which are recorded at a sampling frequency of 44.1 kHz. After downsampling to 16 kHz, speech-only segments are manually extracted from the recordings.

MoSpeeDi database. We consider French recordings from 20 PD and ALS patients (14 males, 6 females) and 20 healthy speakers (10 males, 10 females) from Geneva University Hospitals and University of Geneva. Each speaker utters 54 pseudo-words, which are recorded at a sampling frequency of 44.1 kHz. After downsampling to 16 kHz, speech-only segments are extracted from the recordings using an energy-based voice activity detector [24].

Table 2: *Architecture of the proposed CNN-based classifier operating on pairwise distance matrices.*

Layer	Description
Input	Size: (1xSxS) input distance matrix
Conv2d + Relu	Channel: in=1, out=16, Filter: 10x10, Stride: 1
Maxpool2d	Channel: in=16, out=16, Filter: 2x2, Stride: 2
Conv2d + Relu	Channel: in=16, out=16, Filter: 10x10, Stride: 1
Maxpool2d	Channel: in=16, out=16, Filter: 2x2, Stride: 2
Dropout	Probability: 0.5
FC + Relu	Input: 784, Output: 128
FC + Softmax	Input: 128, Output: 2

3.2. Articulatory posterior representation

AP representations are extracted as in [22], where frame-level posteriors of four articulatory categories are computed, i.e., manner of articulation (e.g., degree of constriction), place of constriction, height of the tongue, and vowel. Posteriors for each category are estimated using CNNs trained on healthy speech data from the AMI corpus [25] based on acoustic phoneme-to-articulatory feature mappings [21]. By concatenating all extracted APs, $F_1 = 53$ features per time frame are obtained. For details on the training procedure for AP feature extraction, the reader is referred to [22].

3.3. Baseline networks

To demonstrate the advantages of the proposed approach, the following two baseline systems B-CNN$_1$ and B-CNN$_2$ are considered.

B-CNN$_1$. We have implemented a baseline CNN adapted from [14], which is trained on log magnitude of STFT representations of short (i.e., 160 ms) segments of speech with 50% overlap. The STFT representations are computed using 10 ms Hanning windows without overlap, resulting in 129 frequency bins for each time frame. The final decision for an unseen speaker is made by applying soft voting on the segment-level CNN prediction scores. To demonstrate the advantage of using AP representations instead of STFT, such a baseline CNN is also trained on the logarithm of AP representations. The architecture of this baseline system is summarized in Table 3.

B-CNN$_2$. To further establish the advantages of the proposed end-to-end CNN framework (which uses a front-end feature extraction layer), a second baseline is implemented where the proposed CNN-based classifier in Section 2.3 is trained on distance matrices computed directly from AP representations (i.e., without using the front-end feature extraction layer). To compute such distance matrices, Kullback-Leibler divergence is used as the local distance measure in (1). The architecture of this baseline system is the same as in Table 2.

3.4. Training and evaluation

The validation strategy on the PC-GITA and MoSpeeDi databases is a stratified speaker-independent 10-fold and 5-fold cross-validation framework, respectively (i.e., speakers in each fold are different). In each training fold, a development fold with the same size as the test fold is set aside for early-stopping. Z-score normalization is applied to all input representations. All networks are trained using the stochastic gradient descent (SGD) algorithm and the cross-entropy loss. The batch size is 256, and the initial learning rate is 0.05. The

Table 3: *Architecture of the baseline B-CNN*$_1$ *adapted from [14].*

Layer	Description
Input	Size: (1xFx16); F: dimension of input representation
Conv1d + Relu	Channel: in=1, out=32, Filter: Fx1, Stride: 1
Conv1d + Relu	Channel: in=32, out=16, Filter: 1x4, Stride: 1
Dropout	Probability: 0.5
FC + Relu	Input: 208, Output: 128
FC + Softmax	Input: 128, Output: 2

learning rate is divided by 5 each time the loss on the development set does not decrease for 5 consecutive iterations. The training is stopped either after 100 epochs or after the learning rate has reached the value 10^{-6}.

Random weight initialization is used for the baselines B-CNN$_1$ and B-CNN$_2$. The weights on the first convolution layer of the trained baseline B-CNN$_1$ are used to initialize the front-end feature extraction layer of the proposed end-to-end CNN. The weights of the trained baseline B-CNN$_2$ are used to initialize the classifier layers of the proposed end-to-end CNN.

The number of total samples (training/testing) available for the different considered networks is as follows. Using the STFT representation for B-CNN$_1$ results in 17383 (PC-GITA) and 25197 (MoSpeeDi) segments. Using the AP representation for B-CNN$_1$ results in 17368 (PC-GITA) and 25907 (MoSpeeDi) segments. The number of distance matrices computed from all pairs of reference and test AP representations for B-CNN$_2$ and the proposed CNN is 96000 (PC-GITA) and 25920 (MoSpeeDi).

The dysarthric speech detection performance is evaluated in terms of the area under ROC curve (AUC). In addition, we also compute the classification accuracy using a decision threshold of 0.5. To reduce the impact of the random seed on the final model parameters, we have trained all networks with 3 different random seeds. The reported performance measures are the mean and standard deviation of the performance obtained by models trained using different seeds.

3.5. Results

Table 4 presents the AUC and classification accuracy values obtained using B-CNN$_1$ on STFT and AP representations for both considered databases. It can be observed that the AP representation yields a better performance than the STFT on both databases, with a particularly significant improvement observed for the PC-GITA database. These results are to be expected given the advantages of articulatory modeling of speech using AP as described in Section 1.

It should be noted that the CNN proposed in [14] was trained on the PC-GITA database using speech segments centered at transitions between voiced and unvoiced regions. However, although not presented here due to space constraints, using such segments did not result in a better performance than the performance presented in Table 4. Further, it should be noted that [14] uses more recordings than the word recordings we have used here. To ensure that the conclusions derived in this paper on the advantages of the proposed approach as opposed to B-CNN$_1$ are still valid even when more recordings are available for use in B-CNN$_1$, we have investigated the performance of B-CNN$_1$ using AP representations on both databases when all available recordings are used (rather than just words).

Using all available recordings and the AP representation for B-CNN$_1$ results in 74762 (PC-GITA) and 54626 total available segments. In this case, B-CNN$_1$ yields AUC and accuracy values of 0.78 and 73.33% on the PC-GITA database and 0.75 and 60.00% on the MoSpeeDi database. When comparing these results to the ones obtained using only word recordings (cf. entries for AP representations in Table 4), we observe that increasing the used speech material does not significantly improve the dysarthric speech detection performance of B-CNN$_1$. In summary, the presented results demonstrate the advantage of using AP representations as opposed to the STFT representations used in [14]. In the following, the performance of both baseline systems B-CNN$_1$ and B-CNN$_2$ and of the proposed end-to-end CNN is compared when AP representations are used.

Table 4: *Mean and standard deviation of the AUC score and classification accuracy [%] using the baseline B-CNN$_1$ with STFT and AP representations on the PC-GITA and MoSpeeDi databases.*

Database	Input representation	AUC	Accuracy
Spanish PC-GITA	STFT	0.56 ± 0.03	53.67 ± 3.29
Spanish PC-GITA	AP	0.75 ± 0.00	72.00 ± 0.81
French MoSpeeDi	STFT	0.64 ± 0.02	52.50 ± 0.00
French MoSpeeDi	AP	0.73 ± 0.03	60.83 ± 3.12

Table 5: *Mean and standard deviation of the AUC score and classification accuracy [%] using the baseline B-CNN$_1$ and B-CNN$_2$ and the proposed pairwise distance-based approach with a front-end feature extraction layer on the PC-GITA and MoSpeeDi databases.*

Database	CNN	AUC	Accuracy
Spanish PC-GITA	Baseline B-CNN$_1$	0.75 ± 0.00	72.00 ± 0.81
Spanish PC-GITA	Baseline B-CNN$_2$	0.78 ± 0.01	68.33 ± 0.74
Spanish PC-GITA	Proposed	$\mathbf{0.83} \pm 0.01$	$\mathbf{77.67} \pm 0.47$
French MoSpeeDi	Baseline B-CNN$_1$	0.73 ± 0.03	60.83 ± 3.11
French MoSpeeDi	Baseline B-CNN$_2$	0.77 ± 0.00	70.83 ± 2.35
French MoSpeeDi	Proposed	$\mathbf{0.84} \pm 0.02$	$\mathbf{76.67} \pm 4.25$

Table 5 presents the AUC and classification accuracy values of the baseline systems B-CNN$_1$ and B-CNN$_2$ and of the proposed approach on both databases. Bold entries indicate the maximum performance for each database. It can be observed that the proposed pairwise distance-based CNN with front-end feature extraction outperforms both considered baselines in terms of both performance measures on both databases. Comparing the difference in performance between the proposed framework and B-CNN$_2$ shows that incorporating a feature extraction front-end significantly improves the performance in comparison to computing distance matrices directly on AP representations. Analyzing the learned representations from the feature extraction front-end remains a topic for future investigation.

In summary, the presented results show that the proposed pairwise distance-based CNN with a front-end feature extraction layer is successfully applicable to the dysarthric speech detection task. Although a small number of utterances per speaker are used, the proposed approach outperforms baseline systems for different databases with different languages and disorders.

4. CONCLUSION

In this study, we explored the feasibility of automatic dysarthric speech detection using a pairwise distance-based CNN. The proposed approach compares frame-level distance patterns between phonetically-balanced AP representations from healthy (i.e., reference) and test speakers. After extracting features from such representations and processing their distance matrix, a CNN-based classifier predicts whether the test representation is from a healthy or dysarthric speaker. Feature extraction, distance matrix computation, and classification are jointly optimized in an end-to-end framework. Experimental results on two dysarthric speech databases have shown that the proposed approach is generalizable across languages, obtaining a high dysarthric speech detection accuracy and outperforming state-of-the-art CNN-based systems.

5. REFERENCES

[1] F. L. Darley, A. E. Aronson, and J. R. Brown, “Differential diagnostic patterns of dysarthria,” *Journal of Speech, Language, and Hearing Research*, vol. 12, no. 2, pp. 246–269, Jun. 1969.

[2] L. Gavidia-Ceballos and J. Hansen, “Direct speech feature estimation using an iterative EM algorithm for vocal fold pathology detection,” *IEEE Transactions on Biomedical Engineering*, vol. 43, no. 4, pp. 373–383, Apr. 1996.

[3] L. Baghai-Ravary and S. Beet, *Automatic speech signal analysis for clinical diagnosis and assessment of speech disorders*. New York, USA: Springer, Aug. 2012.

[4] S. Hegde, S. Shetty, S. Rai, and T. Dodderi, “A survey on machine learning approaches for automatic detection of voice disorders,” *Journal of Voice*, vol. 33, no. 6, pp. 947.e11–947.e33, Nov. 2019.

[5] J. Gómez-García, L. Moro-Velázquez, and J. Godino-Llorente, “On the design of automatic voice condition analysis systems. part i: Review of concepts and an insight to the state of the art,” *Biomedical Signal Processing and Control*, vol. 51, pp. 181–199, May 2019.

[6] 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, May 2012.

[7] J. R. Orozco-Arroyave, F. Hönig, J. Arias-Londoño, J. Bonilla, S. Skodda, J. Rusz, and E. Nöth, “Voiced/unvoiced transitions in speech as a potential bio-marker to detect Parkinson’s disease,” in *Proc. Annual Conference of the International Speech Communication Association*, Dresden, Germany, Sep. 2015, pp. 95–99.

[8] D. Hemmerling, J. R. Orozco-Arroyave, A. Skalski, J. Gajda, and E. Nöth, “Automatic detection of Parkinson’s disease based on modulated vowels,” in *Proc. Annual Conference of the International Speech Communication Association*, San Francisco, USA, Sep. 2016, pp. 1190–1194.

[9] 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, Feb. 2010.

[10] I. Kodrasi and H. Bourlard, “Super-Gaussianity of speech spectral coefficients as a potential biomarker for dysarthric speech detection,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Brighton, UK, May 2019, pp. 6400–6404.

[11] ——, “Spectro-temporal sparsity characterization for dysarthric speech detection,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 28, no. 1, pp. 1210–1222, Dec. 2020.

[12] I. Kodrasi, M. Pernon, M. Laganro, and H. Bourlard, “Automatic discrimination of apraxia of speech and dysarthria using a minimalistic set of handcrafted features,” in *Proc. Annual Conference of the International Speech Communication Association*, Shanghai, China, Oct. 2020.

[13] A. Hernandez, E. J. Yeo, S. Kim, and M. Chung, “Dysarthria Detection and Severity Assessment Using Rhythm-Based Metrics,” in *Proc. 21st Annual Conference of the International Speech Communication Association*, Shanghai, China, Sep. 2020, pp. 2897–2901.

[14] J. Vasquez, J. R. Orozco, and E. Noeth, “Convolutional neural network to model articulation impairments in patients with parkinson’s disease,” in *In Proc. Annual Conference of the International Speech Communication Association*, Stockholm, Sweden, Aug. 2017, pp. 314–318.

[15] E. Vaiciukynas, A. Gelzinis, A. Verikas, and M. Bacauskiene, “Parkinson’s disease detection from speech using convolutional neural networks,” in *In Proc. International Conference on Smart Objects and Technologies for Social Good*. Pisa, Italy: Springer International Publishing, Nov. 2017, pp. 206–215.

[16] K. An, M. Kim, K. Teplansky, J. Green, T. Campbell, Y. Yunusova, D. Heitzman, and J. Wang, “Automatic early detection of amyotrophic lateral sclerosis from intelligible speech using convolutional neural networks,” in *Proc. Annual Conference of the International Speech Communication Association*, Hyderabad, India, Sep. 2018.

[17] S. Bhati, L. M. Velazquez, J. Villalba, and N. Dehak, “LSTM siamese network for parkinson’s disease detection from speech,” in *In Proc. IEEE Global Conference on Signal and Information Processing*, Ottawa, Canada, Nov. 2019, pp. 1–5.

[18] J. Mallela, A. Illa, Y. Belur, N. Atchayaram, R. Yadav, P. Reddy, D. Gope, and P. K. Ghosh, “Raw Speech Waveform Based Classification of Patients with ALS, Parkinson’s Disease and Healthy Controls Using CNN-BLSTM,” in *Proc. 21st Annual Conference of the International Speech Communication Association*, Shanghai, China, Sep. 2020, pp. 4586–4590.

[19] N. Cummins, A. Baird, and B. W. Schuller, “Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning,” *Methods*, vol. 151, pp. 41–54, Dec. 2018, health Informatics and Translational Data Analytics.

[20] D. Ram, L. Miculicich, and H. Bourlard, “Neural network based end-to-end query by example spoken term detection,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, no. 1, pp. 1416–1427, Apr. 2020.

[21] R. Rasipuram and M. Magimai.-Doss, “Articulatory feature based continuous speech recognition using probabilistic lexical modeling,” *Computer Speech & Language*, vol. 36, pp. 233–259, Mar. 2016.

[22] S. P. Dubagunta and M. Magimai-Doss, “Using Speech Production Knowledge for Raw Waveform Modelling Based Styrian Dialect Identification,” in *In Proc. Annual Conference of the International Speech Communication Association*, Graz, Austria, Sep. 2019, pp. 2383–2387.

[23] J. R. Orozco, J. D. Arias-Londoño, J. Vargas-Bonilla, M. González-Rátiva, and E. Noeth, “New spanish speech corpus database for the analysis of people suffering from parkinson’s disease,” in *Proc. 9th International Conference on Language Resources and Evaluation*, Reykjavik, Iceland, May. 2014, pp. 342–347.

[24] P. Boersma, “PRAAT, a system for doing phonetics by computer,” *Glot International*, vol. 5, no. 9, pp. 341–345, Jan. 2002.

[25] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, “The ami meeting corpus: A pre-announcement,” in *In Proc. 2nd International Conference on Machine Learning for Multimodal Interaction*, ser. MLMI’05. Berlin, Heidelberg: Springer-Verlag, July 2005, pp. 28–39.