Temporal Envelope and Fine Structure Cues for Dysarthric Speech Detection Using CNNs

Ina Kodrasi, *Senior Member, IEEE*

***Abstract*—Deep learning-based techniques for automatic dysarthric speech detection have recently attracted interest in the research community. State-of-the-art techniques typically learn neurotypical and dysarthric discriminative representations by processing time-frequency input representations such as the magnitude spectrum of the short-time Fourier transform (STFT). Although these techniques are expected to leverage perceptual dysarthric cues, representations such as the magnitude spectrum of the STFT do not necessarily convey perceptual aspects of complex sounds. Inspired by the temporal processing mechanisms of the human auditory system, in this paper we factor signals into the product of a slowly varying envelope and a rapidly varying fine structure. Separately exploiting the different perceptual cues present in the envelope (i.e., phonetic information, stress, and voicing) and fine structure (i.e., pitch, vowel quality, and breathiness), two discriminative representations are learned through a convolutional neural network and used for automatic dysarthric speech detection. Experimental results show that processing both the envelope and fine structure representations yields a considerably better dysarthric speech detection performance than processing only the envelope, fine structure, or magnitude spectrum of the STFT representation.**

***Index Terms*—temporal envelope, temporal fine structure, dysarthria, Parkinson's disease, convolutional neural network**

I. Introduction

Neurological disorders such as Parkinson's disease (PD) can cause dysarthria, resulting in disrupted speech production across different dimensions. To detect and manage dysarthria, clinicians exploit perceptual assessments typically involving evaluation by ear of clinical-perceptual signs of dysarthric speech, e.g., articulation deficiencies, vowel quality changes, pitch variation, breathiness, or rhythm disruptions [1]. These perceptual evaluations are subject to the expertise of the clinician and can be time-consuming [2]. To complement the perceptual assessment of clinicians, objective dysarthric speech processing techniques have been proposed. Such techniques can assist clinicians by automatically detecting the presence of dysarthria [3]–[5] or by automatically evaluating the patient's intelligibility and dysarthria severity [6]–[9].

Typical automatic dysarthric speech detection techniques are based on handcrafting acoustic features aiming to characterize the clinical-perceptual signs of dysarthria [10]–[18]. Acoustic features such as jitter, shimmer, or fundamental frequency have been used to quantify impacted phonation [10]–[12]. Acoustic features such as Mel frequency cepstral coefficients and spectro-temporal sparsity measures have been used to quantify articulation deficiencies [12]–[16]. Further, the envelope modulation spectrum and durational measures of vocalic and intervocalic segments have been used to quantify rhythm disruptions [17]–[19]. Although successful results have been reported using handcrafted features, such features may fail to characterize more abstract but similarly important perceptual dysarthric cues. Consequently, there has been a growing interest in the research community to develop deep learning-based automatic dysarthric speech detection techniques [20]–[26].

In [20], raw neurotypical and dysarthric speech segments have been used to train a long short-term memory Siamese networks that learns discriminative representations. Raw speech segments have also been used in [21], where convolutional neural networks (CNNs) have been trained instead. Given the limited amount of pathological training data, contributions exploiting raw speech segments are seldom. Instead, mainstream techniques rely on processing the magnitude spectrum of time-frequency representations such as the Mel spectogram [23]–[25], the continuous wavelet transform [22], or the short-time Fourier transform (STFT) [22], [23], [26]. Although these techniques are expected to leverage perceptual dysarthric cues, such representations do not necessarily convey perceptual aspects of complex sounds [27].

Within the cochlea, speech signals are filtered into a series of narrowband signals with a slowly varying envelope imposed on a rapidly oscillating carrier, i.e., the temporal fine structure. The relative importance of the temporal envelope and fine structure to speech perception has been the subject of a wide range of literature for decades, with particular focus on the importance of these cues for speech intelligibility in the presence of interference and the effects of hearing loss on the processing of these cues in the auditory nerve [28]–[31]. Furthermore, processing the temporal envelope and/or fine structure has been crucial for applications such as automatic speech recognition or speech enhancement [32]–[34]. The importance of fine structure cues for dysarthric speech assessment has been recently demonstrated in [35], where these cues have been extracted using a single frequency filtering representation and exploited in an i-vector based dysarthria detection system. Although the relative importance of the temporal envelope and fine structure for speech perception is still debated (cf., [36]), it is established that envelope signals contain phonetic information as well as stress and voicing information, whereas fine structure signals are important for pitch perception and vowel quality [27], [28].

Inspired by these temporal processing mechanisms of the human auditory system, in this paper we propose a deep learning-based dysarthric speech detection technique which

This work was supported by the Swiss National Science Foundation project no CRSII5_173711 on "Motor Speech Disorders: characterizing phonetic speech planning and motor speech programming/execution and their impairments".

I. Kodrasi is with the Idiap Research Institute, Martigny, 1920 Switzerland (e-mail: ina.kodrasi@idiap.ch).

separately processes the temporal envelope and fine structure signals. Two discriminative representations separately learned from the temporal envelope and fine structure using CNNs are then exploited for automatic dysarthric speech detection. To the best of our knowledge, the extraction of temporal fine structure signals through an auditory-inspired filter bank and their use in deep learning-based approaches has never been investigated.

Experimental results in Section IV show that the temporal envelope contains more cues for dysarthric speech detection than the temporal fine structure. Further, it is shown that the proposed approach which exploits cues from both signals to learn two discriminative representations provides a considerable performance increase as opposed to learning a single discriminative representation from inputs where dysarthric cues are partially lost or intermingled (such as in the magnitude spectrum of the STFT representation).

II. Temporal Envelope and Fine Structure Dysarthric Speech Detection

In the following, the proposed temporal envelope and fine structure (TEFS)-based dysarthric speech detection system is described. Section II-A presents the computation of the temporal envelope and fine structure representations, whereas Section II-B presents the used CNN.

A. Temporal envelope and fine structure representations

We denote the speech signal of a neurotypical or dysarthric speaker by $s(n)$, with n being the time index. When a clinician listens to this signal to conduct their perceptual assessment, the cochlea processes the signal through frequency analysis and temporal envelope and fine structure decomposition. To mimic cochlear frequency analysis, we use a bank of K band-pass filters to split the signal $s(n)$ into K complementary frequency bands of equal width along the human basiliar membrane [28]. Let $s_c(k,n)$ denote the subband signal at the output of the k-th band-pass filter, with $k = 1,\ldots,K$. The subband temporal envelope and fine structure signals are computed through the analytic representation of $s_c(k,n)$, i.e.,

$$s_a(k,n) = s_c(k,n) + j\mathcal{H}\{s_c(k,n)\}, \tag{1}$$

where $\mathcal{H}\{s_c(k,n)\}$ denotes the Hilbert transform of $s_c(k,n)$. Based on (1), the subband temporal envelope and fine structure signals $e_c(k,n)$ and $f_c(k,n)$ can be computed as

$$e_c(k,n) = \sqrt{s_c^2(k,n) + \mathcal{H}^2\{s_c(k,n)\}}, \tag{2}$$

$$f_c(k,n) = \cos\left[\arctan\left(\frac{\mathcal{H}\{s_c(k,n)\}}{s_c(k,n)}\right)\right]. \tag{3}$$

These signals are then averaged within time frames of length L_w to create the temporal envelope and fine structure representations $E_c(k,l)$ and $F_c(k,l)$, $l = 1,\ldots,L$, with L being the total number of time frames in $s(n)$. To further emphasize these representations, log scaling is applied to obtain the final envelope and fine structure representations $E(k,l)$ and $F(k,l)$. Since $E_c(k,l) > 0$ (cf. (2)), the final envelope representation is obtained as $E(k,l) = \log_{10} E_c(k,l)$. Since $-1 \leq F_c(k,l) \leq 1$ (cf. (3)), the final fine structure representation is obtained as $F(k,l) = \text{sgn}\{F_c(k,l)\}\log_{10}|F_c(k,l)|$ such that zero-crossings are preserved.

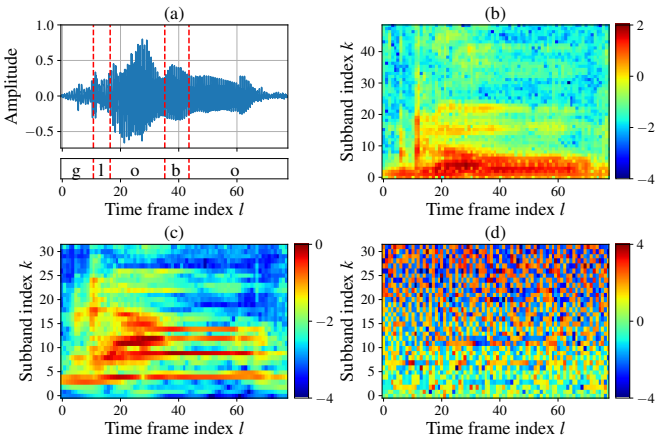

Fig. 1. Different representations of the exemplary utterance *globo*: (a) time domain signal $s(n)$, (b) magnitude spectrum of the STFT using $L_w = 6$ ms, (c) envelope $E(k,l)$ using $K = 32$ and $L_w = 6$ms, and (d) fine structure $F(k,l)$ using $K = 32$ and $L_w = 6$ms.

Fig. 1(a) depicts an exemplary utterance $s(n)$ from the database described in Section III-A. The temporal envelope and fine structure representations $E(k,l)$ and $F(k,l)$ for this utterance computed using $K = 32$ and $L_w = 6$ ms are depicted in Figs. 1(c) and 1(d). These representations convey different perceptual cues, with the envelope representation conveying phonetic information as well as stress and voicing information and the fine structure representation conveying pitch and vowel quality information. For completeness, the commonly used (logarithm of the) magnitude spectrum of the STFT representation of $s(n)$ using $L_w = 6$ ms is depicted in Fig. 1(b), where these different perceptual cues are either partially lost or intermingled.[1] This perceptual information loss occurs not only because the phase of the STFT is disregarded, but also because the STFT uses uniform filter banks which do not approximate well auditory frequency analysis.

B. Convolutional neural network

Once a signal representation is computed, the CNN depicted in the block diagram in Fig. 2 can be trained for automatic dysarthric speech detection as in [22], [26]. The CNN receives as input $(K \times B)$–dimensional neurotypical and dysarthric speech representations (envelope, fine structure, STFT, or any other time-frequency representation), with B being a user-defined number of time frames. Through alternating between convolutional and pooling layers, the CNN is expected to extract robust discriminative representations of neurotypical and dysarthric speech. These extracted representations are then exploited in fully-connected layers (FCLs) trained to decide whether the $(K \times B)$–dimensional input representation corresponds to a neurotypical or dysarthric speaker. While this

[1] It should be noted that the STFT representation results in a trade-off between spectral and temporal resolution. Hence, although the same window length $L_w = 6$ ms is used to compute the STFT, the number of STFT subbands differs from the number of subbands used in the envelope and fine structure representations.

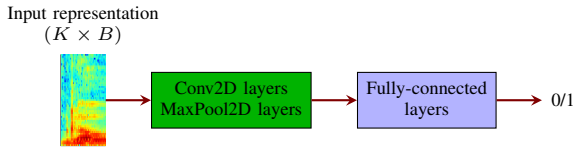

Fig. 2. Block diagram of the baseline CNN-based dysarthric speech detection system from [22].

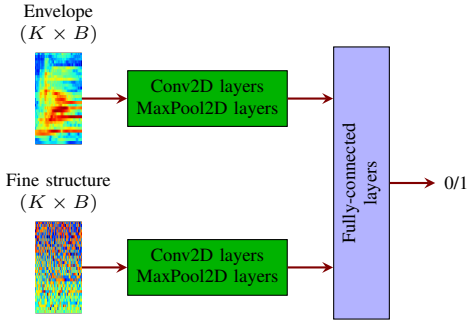

Fig. 3. Block diagram of the proposed temporal envelope and fine structure-based dysarthric speech detection system.

approach can be used on the individual envelope and fine structure representations described in Section II-A, it is sub-optimal since only the cues available in one representation would be exploited (cf. Section IV).

To exploit cues available in both the temporal envelope and fine structure representations, we propose to use the TEFS-based dysarthric speech detection system depicted in Fig. 3. As shown in this figure, we use individual convolutional and pooling layers that operate on the envelope and fine structure representations. Two discriminative representations are extracted and jointly exploited in FCLs trained to detect dysarthric speech. As shown in Section IV, such an approach yields a considerably better performance than using the system depicted in Fig. 2 on the individual envelope, fine structure, or magnitude of the STFT representations.

III. Material and Method

A. Database

We consider Spanish recordings of 50 PD patients and 50 neurotypical speakers from the PC-GITA database [37]. The database is well balanced in terms of age and gender and the recordings are captured in a sound proof booth at a sampling frequency of 44.1 kHz. For the results presented in the following, we use recordings of 24 different words and of a phonetically balanced text downsampled to 16 kHz. The average length of the available speech material for each speaker is 32.1 s.

B. Proposed TEFS-based network

The proposed TEFS-based dysarthric speech detection system depicted in Fig. 3 operates on segments of envelope and fine structure representations. For the results presented in this paper, these segments are computed as follows.

TABLE I
Architecture of the proposed TEFS-based dysarthric speech detection technique. BN refers to batch normalization.

Layers	Envelope or fine structure branch
Input	$(K \times B)$-dimensional envelope
Conv2D+ReLU+BN	in=1, out=64, kernel=$(2,2)$, stride=$(1,1)$
MaxPool2D	in=64, out=64, kernel=$(2,2)$, stride=$(2,2)$
Conv2D+ReLU+BN	in=64, out=64, kernel=$(3,3)$, stride=$(1,1)$
MaxPool2D	in=64, out=64, kernel=$(2,2)$, stride=$(2,2)$
Dropout	probability = 0.5
FCL+ReLU	in=8448, out=128
FCL+Softmax	in=128, out=2

We design band-pass filters spanning the range from 80 Hz to 7200 Hz, with cut-off frequencies spaced in equal steps along the cochlear frequency map [28], [38]. The number of filters used is $K = 32$. After band-pass filtering the input signal, the envelope and fine structure representations are computed as described in Section II-A using $L_w = 6$ ms. Finally, $(K \times B)$-dimensional segments using $B = 50$ and a 50% overlap are extracted and used as inputs to the system. Table I summarizes the architecture of the proposed system, which has approximately 1 million trainable parameters. As shown in this table, the same architecture (adapted from [22]) is used for both the envelope and fine structure branches.

C. Baseline networks

To analyze the individual cues available for dysarthric speech detection in the envelope and fine structure representations, the baseline CNN depicted in Fig. 2 is separately trained on the envelope and fine structure representations computed as described in Section III-B. To further demonstrate the advantages of the proposed approach, we have also trained such a baseline CNN on the magnitude of the STFT representation.

The STFT is computed using a weighted overlap-add framework with a Hanning analysis window without overlap. As previously mentioned, the STFT yields a trade-off between spectral and temporal resolution. For a fair comparison, we consider an STFT analysis window length $L_w = 3.875$ ms, such that the same spectral dimension (i.e., $K = 32$) is obtained as for the envelope and fine structure representations. After computing the STFT, $(K \times B)$-dimensional segments using $B = 50$ and a 50% overlap are extracted and used as inputs to the system.

Two architectures A_1 and A_2 are considered for these baseline networks. For A_1, the same architecture as the one presented in Table I for the individual envelope or fine structure branches is used. After the dropout layer, a FCL (with an input dimension of 4224 and output dimension of 2) followed by the softmax function is used. Such an architecture has approximately 45 thousand trainable parameters. Since A_1 has a considerably lower number of parameters than the proposed system in Table I, we also consider the deeper architecture A_2 shown in Table II. This architecture has approximately 1 million trainable parameters, comparable to the proposed system in Table I (cf. Section III-B).

TABLE II
ARCHITECTURE A_2 FOR THE BASELINE SYSTEMS. BN REFERS TO BATCH NORMALIZATION.

Layers	
Input	$(K \times B)$-dimensional envelope
Conv2D+ReLU+BN	in=1, out=64, kernel=$(2,2)$, stride=$(1,1)$
MaxPool2D	in=64, out=64, kernel=$(2,2)$, stride=$(2,2)$
Conv2D+ReLU+BN	in=64, out=64, kernel=$(3,3)$, stride=$(1,1)$
MaxPool2D	in=64, out=64, kernel=$(2,2)$, stride=$(2,2)$
Conv2D+ReLU+BN	in=64, out=64, kernel=$(4,4)$, stride=$(1,1)$
MaxPool2D	in=64, out=64, kernel=$(2,2)$, stride=$(2,2)$
Dropout	probability = 0.5
FCL+ReLU	in=256, out=4096
FCL+Softmax	in=4096, out=2

D. Training and evaluation

The evaluation strategy is a speaker-independent stratified 10-fold cross-validation, ensuring that each fold is balanced in terms of gender and in terms of the number of neurotypical and PD speakers. In each training fold, a development set with the same size as the test set is used for early-stopping. Z-score normalization is applied to all input representations and networks are trained using the stochastic gradient descent algorithm and the cross-entropy loss. The batch size is 128 and the initial learning rate is 0.01. The learning rate is halved if the loss on the development set has not decreased for 5 consecutive iterations. Training is stopped when the learning rate has decreased beyond 10^{-6} or after 100 epochs. The trained models output a prediction score for each of the $(K \times B)$–dimensional segments and the final decision for an unseen speaker is made by applying soft voting on these segment-level prediction scores.

The baseline CNNs trained on the envelope, fine structure, or STFT representations are randomly initialized. The convolutional layers of the proposed TEFS-based technique are initialized with the convolutional layers of trained baseline systems, with the upper branch network in Fig. 3 initialized with the baseline architecture A_1 trained on the envelope representation and the lower branch network in Fig. 3 initialized with the baseline architecture A_1 trained on the fine structure representation.

Dysarthric speech detection performance is evaluated in terms of the area under ROC curve (AUC) and classification accuracy for a decision threshold of 0.5. To reduce the impact of initialization on the final model parameters, we have trained all networks with 5 different random seeds. To reduce the impact of the speaker split into training and testing folds, we have repeated this training procedure for 5 different splits of speakers. Hence, we have trained 250 models for each considered network, i.e., 5 models for each of the 10 folds obtained using 5 different fold splits. The reported performance measures are the mean and standard deviation of the performance obtained across these different models.

IV. RESULTS

Table III presents the performance obtained using the baseline CNNs trained on different input representations and using the proposed TEFS-based technique.

TABLE III
PERFORMANCE USING THE BASELINE SYSTEMS TRAINED ON THE STFT, ENVELOPE, AND FINE STRUCTURE REPRESENTATIONS AND USING THE PROPOSED TEFS-BASED TECHNIQUE.

Network	AUC	Accuracy [%]
A_1 - Magnitude of STFT	0.76 ± 0.14	69.52 ± 14.04
A_2 - Magnitude of STFT	0.79 ± 0.14	69.76 ± 13.71
A_1 - Envelope	0.83 ± 0.14	73.80 ± 11.75
A_2 - Envelope	0.81 ± 0.13	70.50 ± 11.42
A_1 - Fine structure	0.72 ± 0.15	65.68 ± 12.38
A_2 - Fine structure	0.66 ± 0.15	61.40 ± 13.36
TEFS	**0.93 ± 0.08**	**85.72 ± 10.38**

It can be observed that using A_2 for any of the baseline systems typically yields a lower performance than using A_1. Such a result can be explained by the considerably larger number of parameters in A_2 in comparison to A_1, resulting in overfitting and poor generalization performance for A_2. Further, it can be observed that out of the considered baseline systems, using the envelope representation outperforms using the magnitude spectrum of the STFT or the fine structure representation. These results show that the envelope of a signal contains more cues for dysarthric speech detection than its fine structure. Further, these results confirm the advantages of using these auditory-inspired representations for CNN-based dysarthric speech detection.

Finally, Table III shows that the proposed TEFS-based technique yields a better performance than all considered baseline systems for both performance measures, with an AUC of 0.93 an accuracy score of 85.72%. These results show that although dysarthric cues can be more prominent in the envelope than in the fine structure, exploiting both representations is very beneficial for deep learning-based dysarthric speech detection.

V. CONCLUSION

In this paper we have proposed a deep learning-based dysarthric speech detection technique inspired by the temporal processing mechanisms of the human auditory system. The proposed technique relies on decomposing speech signals into their envelope and fine structure counterparts, each containing different perceptual cues for dysarthric speech detection. By separately processing the envelope and fine structure through individual convolutional and pooling layers, two discriminative representations are learned and jointly exploited for dysarthric speech detection. Experimental results on a Spanish database of neurotypical and PD speakers have shown that the envelope representation contains more discriminative cues than the fine structure representation. Further, experimental results have shown that exploiting both envelope and fine structure representations yields a considerably better dysarthric speech detection performance than exploiting only the envelope, fine structure, or STFT representation. In the future, we plan to investigate how the incorporation of more complex auditory models affects the extracted discriminative representations and the final performance of the system. Further, we plan to investigate different architectures for processing the temporal envelope and fine structure cues.

REFERENCES

[1] J. R. Duffy, *Motor speech disorders*. St Louis, USA: Elsevier Mosby, 2019.

[2] S. Fonville, H. B. van der Worp, P. Maat, M. Aldenhoven, A. Algra, and J. van Gijn, "Accuracy and inter-observer variation in the classification of dysarthria from speech recordings," *IEEE Transactions on Speech and Audio Processing*, vol. 255, no. 10, pp. 1545–1548, Oct. 2008.

[3] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, and L. O. Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 4, pp. 1015–1022, Apr. 2009.

[4] S. Bhati, L. M. Velazquez, J. Villalba, and N. Dehak, "LSTM Siamese network for Parkinson's disease detection from speech," in *Proc. IEEE Global Conference on Signal and Information Processing*, Ottawa, Canada, Nov. 2019, pp. 1–5.

[5] P. Janbakhshi, I. Kodrasi, and H. Bourlard, "Subspace-based learning for automatic dysarthric speech detection," *IEEE Signal Processing Letters*, vol. 28, pp. 96–100, 2021.

[6] M. J. Kim, Y. Kim, and H. Kim, "Automatic intelligibility assessment of dysarthric speech using phonologically-structured sparse linear model," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 694–704, Feb. 2015.

[7] I. Laaridh, W. B. Kheder, C. Fredouille, and C. Meunier, "Automatic prediction of speech evaluation metrics for dysarthric speech," in *Proc. Annual Conference of the International Speech Communication Association*, Stockholm, Sweden, Aug. 2017, pp. 1834–1838.

[8] C. Bhat, B. Vachhani, and S. K. Kopparapu, "Automatic assessment of dysarthria severity level using audio descriptors," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, New Orleans, USA, Mar. 2017, pp. 5070–5074.

[9] P. Janbakhshi, I. Kodrasi, and H. Bourlard, "Automatic pathological speech intelligibility assessment exploiting subspace-based analyses," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1717–1728, May 2020.

[10] 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.

[11] D. Sztahó, G. Kiss, and K. Vicsi, "Estimating the severity of Parkinson's disease from speech using linear regression and database partitioning," in *Proc. Annual Conference of the International Speech Communication Association*, Dresden, Germany, Sep. 2015, pp. 498–502.

[12] 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.

[13] 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.

[14] 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.

[15] ——, "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.

[16] I. Kodrasi, M. Pernon, M. Laganaro, 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, pp. 4991–4995.

[17] J. Liss, S. LeGendre, and A. Lotto, "Discriminating dysarthria type from envelope modulation spectra," *Journal of Speech, Language, and Hearing research*, vol. 53, no. 5, pp. 1246–55, Oct. 2010.

[18] A. Hernandez, E. J. Yeo, S. Kim, and M. Chung, "Dysarthria detection and severity assessment using rhythm-based metrics," in *Proc. Annual Conference of the International Speech Communication Association*, Shanghai, China, Sep. 2020, pp. 2897–2901.

[19] T. H. Falk, R. Hummel, and W.-Y. Chan, "Quantifying perturbations in temporal dynamics for automated assessment of spastic dysarthric speech intelligibility," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Prague, Czech Republic, 2011, pp. 4480–4483.

[20] J. M. N. Zeghidour, "Learning to detect dysarthria from raw speech," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Brighton, UK, May 2019, pp. 5831–5835.

[21] 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. Annual Conference of the International Speech Communication Association*, Shanghai, China, Sep. 2020, pp. 4586–4590.

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

[23] E. Vaiciukynas, A. Gelzinis, A. Verikas, and M. Bacauskiene, "Parkinson's disease detection from speech using convolutional neural networks," in *Proc. International Conference on Smart Objects and Technologies for Social Good*, Pisa, Italy, Nov. 2017, pp. 206–215.

[24] 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.

[25] J. Vasquez-Correa, T. Arias-Vergara, M. Schuster, J. Orozco-Arroyave, and E. Noeth, "Parallel representation learning for the classification of pathological speech: Studies on Parkinson's disease and cleft lip and palate," *Speech Communication*, vol. 122, pp. 56–67, Sep. 2020.

[26] P. Janbakhshi, I. Kodrasi, and H. Bourlard, "Automatic dysarthric speech detection exploiting pairwise distance-based convolutional neural networks," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toronto, Canada, May 2021, pp. 7328–7332.

[27] S. Rosen, R. P. Carlyon, C. J. Darwin, and I. J. Russell, "Temporal information in speech: acoustic, auditory and linguistic aspects," *Philosophical Transactions of the Royal Society of London*, vol. 336, no. 1278, pp. 367–373, Jun. 1992.

[28] Z. M. Smith, B. Delgutte, and A. J. Oxenham, "Chimaeric sounds reveal dichotomies in auditory perception," *Nature*, vol. 416, no. 6876, pp. 87–90, Mar. 2002.

[29] S. Liu and F. Zeng, "Temporal properties in clear speech perception," *Journal of the Acoustical Society of America*, vol. 120, no. 1, pp. 424–432, Jun. 2006.

[30] K. Hopkins and B. Moore, "The importance of temporal fine structure for the intelligibility of speech in complex backgrounds," *in Proc. International Symposium on Auditory and Audiological Research 3*, vol. 125, pp. 442–446, Feb. 2009.

[31] K. Henry, S. Kale, and M. Heinz, "Noise-induced hearing loss increases the temporal precision of complex envelope coding by auditory-nerve fibers," *Frontiers in Systems Neuroscience*, vol. 8, no. 20, Feb. 2014.

[32] N. Moritz, J. Anemueller, and B. Kollmeier, "An auditory inspired amplitude modulation filter bank for robust feature extraction in automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 1926–1937, Nov. 2015.

[33] A. Purushothaman, A. Sreeram, R. Kumar, and S. Ganapathy, "Deep learning based dereverberation of temporal envelopes for robust speech recognition," in *Proc. Annual Conference of the International Speech Communication Association*, Shanghai, China, Oct. 2020, pp. 1688–1692.

[34] I. Thoidis, L. Vrysis, D. Markou, and G. Papanikolaou, "Temporal auditory coding features for causal speech enhancement," *Electronics*, vol. 9, no. 10, pp. 1698–1715, Oct. 2020.

[35] K. Gurugubelli and A. K. Vuppala, "Analytic phase features for dysarthric speech detection and intelligibility assessment," *Speech Communication*, vol. 121, pp. 1–15, Aug. 2020.

[36] S. Shamma and C. Lorenzi, "On the balance of envelope and temporal fine structure in the encoding of speech in the early auditory system," *Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 2818–2833, May 2013.

[37] 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. International Conference on Language Resources and Evaluation*, Reykjavik, Iceland, May 2014, pp. 342–347.

[38] D. D. Greenwood, "A cochlear frequency-position function for several species–29 years later," *Journal of the Acoustical Society of America*, vol. 87, no. 6, pp. 2592–2605, Jun. 1990.