EXPERIMENTAL INVESTIGATION ON STFT PHASE REPRESENTATIONS FOR DEEP LEARNING-BASED DYSARTHRIC SPEECH DETECTION

Parvaneh Janbakhshi[*,†], *Ina Kodrasi*[*]

[*]Idiap Research Institute, Martigny, Switzerland
[†]École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
{parvaneh.janbakhshi,ina.kodrasi}@idiap.ch

ABSTRACT

Mainstream deep learning-based dysarthric speech detection approaches typically rely on processing the magnitude spectrum of the short-time Fourier transform of input signals, while ignoring the phase spectrum. Although considerable insight about the structure of a signal can be obtained from the magnitude spectrum, the phase spectrum also contains inherent structures which are not immediately apparent due to phase discontinuity. To reveal meaningful phase structures, alternative phase representations such as the modified group delay (MGD) spectrum and the instantaneous frequency (IF) spectrum have been investigated in several applications. The objective of this paper is to investigate the applicability of the unprocessed phase, MGD, and IF spectra for dysarthric speech detection. Experimental results show that dysarthric cues are present in all considered phase representations. Further, it is shown that using phase representations as complementary features to the magnitude spectrum is very beneficial for deep learning-based dysarthric speech detection, with the combination of magnitude and IF spectra yielding a very high performance. The presented results should raise awareness in the research community about the potential of the phase spectrum for dysarthric speech detection and motivate further research into novel architectures that optimally exploit magnitude and phase information.

Index Terms— phase, modified group delay, instantaneous frequency, CNN, dysarthria

1. INTRODUCTION

Dysarthria is a motor speech disorder arising from different conditions of brain damage and manifesting through articulation deficiencies, abnormal speech rhythm, hypernasality, or breathiness [1]. Since dysarthria can be one of the earliest signs of several neurodegenerative disorders, its accurate diagnosis in clinical practice is crucial [2, 3]. The clinical diagnosis of dysarthria is typically done through an auditory-perceptual approach, which can be subjective and time-consuming. To complement the clinical perceptual assessment, automatic dysarthric speech detection approaches have been developed.

Automatic dysarthric speech detection approaches can be broadly categorized into two categories, i.e., approaches based on handcrafted acoustic features combined with classical machine learning algorithms [4–8] and deep learning-based approaches that automatically learn high-level discriminative dysarthric representations [9–14]. Given the potential of deep learning-based approaches to characterize abstract but important acoustic cues beyond the realm of knowledge-based handcrafted features, in this paper we focus on deep learning-based approaches.

Mainstream deep learning-based dysarthric speech detection approaches rely on processing the magnitude spectrum (or features derived from the magnitude spectrum) of time-frequency representations such as the short-time Fourier transform (STFT) or continuous wavelet transform [10–14]. In [10], articulation impairments of patients suffering from dysarthria are modeled through a convolutional neural network (CNN) operating on the magnitude spectrum of the continuous wavelet transform. In [11, 12], a CNN is trained on the STFT magnitude spectrum or Mel frequency cepstral coefficients of neurotypical and dysarthric input signals. The STFT magnitude spectrum is also used in [13, 14] to train unsupervised and supervised auto-encoders for dysarthric speech detection. Although considerable insight about the structure of a signal can indeed be obtained from the magnitude spectrum, there are inherent structures also in the phase spectrum, which however has been largely ignored in automatic dysarthric speech detection techniques.

The disregard of the phase spectrum in speech processing applications arises mainly due to the difficulty in processing the phase and due to the uncertainty about its importance [15, 16]. Since phase is wrapped to its principal value, the phase spectrum is discontinuous. Consequently, the phase spectrum is irregular and does not contain visible spectro-temporal patterns that correlate with our understanding of speech. However, several methods have been developed to derive alternative representations revealing spectro-temporal structures hidden in the phase spectrum. Two such representations are the modified group delay (MGD) spectrum [17, 18] and the instantaneous frequency (IF) spectrum [19, 20]. The MGD and IF spectra reflect the derivative of the phase along the frequency and time axis and have been shown to reveal much more meaningful structures than the unprocessed phase spectrum [20]. In addition, although early studies have demonstrated the unimportance of phase to speech perception [21, 22], more recent studies have established the potential of the phase spectrum in different applications such as speech enhancement [16, 23], automatic speech recognition for neurotypical and dysarthric speech [24, 25], or speech synthesis [26]. The potential of the phase spectrum has also been demonstrated for computational paralinguistic applications such as speaker recognition [27,28] and speech emotion recognition [29, 30].

To the best of our knowledge, the STFT phase spectrum or its alternative representations such as the MGD or IF spectra have never been incorporated in deep learning-based dysarthric speech detection approaches. In a recently proposed approach, we have used the temporal envelope and fine structure (i.e., analytical phase) repre-

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"

sentations of speech signals computed through a Gammatone filter bank and sub-sampling [31]. These input representations are separately processed by CNNs to learn two discriminative representations, which are then jointly used for dysarthric speech detection. Experimental results in [31] show that such an approach yields a considerable performance improvement when compared to processing only the STFT magnitude spectrum. However, it remains unclear whether this substantial performance increase arises because of the incorporation of the analytical phase or because of the auditory-inspired processing through a Gammatone filter bank instead of a uniform STFT filter bank.

In this paper we investigate the applicability of STFT phase representations (i.e., the unprocessed phase spectrum, MGD spectrum, and IF spectrum) for dysarthric speech detection. To this end, we analyze the dysarthric speech detection performance of a CNN which uses only phase representations of input signals. In addition, we analyze whether phase representations provide complementary cues for dysarthric speech detection that cannot be extracted from the magnitude representation. Experimental results show that dysarthric cues are present in all considered phase representations, with the MGD and IF spectra yielding a similar dysarthric speech detection performance as the magnitude spectrum. Further, it is shown that using both the magnitude and any of the phase representations is very beneficial, resulting in a considerable performance improvement as opposed to using a single representation. Finally, it is shown that using the STFT magnitude and IF spectra results in a considerably better performance than using the temporal envelope and fine structure representations from [31].

2. INPUT REPRESENTATIONS

In this section, the computation of input signal representations is presented. Section 2.1 presents the STFT phase representations investigated in this paper. For completeness, Section 2.2 provides a brief review of the temporal envelope and fine structure representations used in [31].

2.1. Short-time Fourier transform magnitude and phase representations

Let us consider the time-domain signal $s(n)$ at time index n. To compute the STFT, the signal is segmented into L segments $s_l(n)$ (with or without overlap) of length N samples. Each segment $s_l(n)$ is multiplied by an analysis window and the discrete Fourier transform is applied to obtain the complex STFT coefficients $S_{k,l}$, $k = 1, \ldots, K$, with K denoting the total number of subbands. The complex STFT coefficients can be expressed as

$$S_{k,l} = |S_{k,l}| e^{\jmath \theta_{k,l}}, \tag{1}$$

with $|S_{k,l}|$ and $\theta_{k,l}$ denoting the magnitude and phase of the l-th segment at the k-th subband. Figs. 1(a) and 1(b) depict the magnitude and phase spectra for an exemplary utterance $s(n)$. It can be observed that while the magnitude spectrum exhibits spectro-temporal patterns where formant and pitch information can be identified, the phase spectrum is irregular and difficult to interpret since the phase is wrapped to its principal value, i.e., $-\pi \leq \theta_{k,l} \leq \pi$.

In this paper we investigate the applicability of two alternative phase representations in deep learning-based dysarthric speech detection which aim to reveal spectro-temporal structures hidden in the phase spectrum, i.e., the modified group delay (MGD) spectrum [17,18] and the instantaneous frequency (IF) spectrum [19,20].

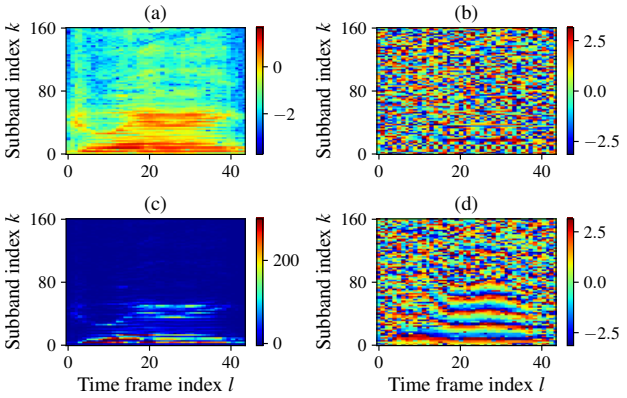

Fig. 1. STFT representations of an exemplary utterance computed using $N = 320$ samples with a 50% overlap and a Hanning analysis window: (a) logarithm of the magnitude, (b) phase, (c) modified group delay, and (d) instantaneous frequency.

The group delay is defined as the negative of the derivative of phase across frequency and can be computed as

$$\tau_{k,l} = \frac{S^{\mathrm{r}}_{k,l} Y^{\mathrm{r}}_{k,l} + S^{\mathrm{i}}_{k,l} Y^{\mathrm{i}}_{k,l}}{|S_{k,l}|^2}, \tag{2}$$

with $Y^{\mathrm{r}}_{k,l}$ and $Y^{\mathrm{i}}_{k,l}$ denoting the real and imaginary part of the STFT coefficients $Y_{k,l}$ of $y_l(n) = n s_l(n)$. To reduce the spiky nature of the group delay spectrum in (2), the MGD spectrum is proposed in [18] which can be computed as

$$\mathrm{MGD}_{k,l} = \mathrm{sign}\left\{ \frac{S^{\mathrm{r}}_{k,l} Y^{\mathrm{r}}_{k,l} + S^{\mathrm{i}}_{k,l} Y^{\mathrm{i}}_{k,l}}{\hat{S}^{2\gamma}_{k,l}} \right\} \left| \frac{S^{\mathrm{r}}_{k,l} Y^{\mathrm{r}}_{k,l} + S^{\mathrm{i}}_{k,l} Y^{\mathrm{i}}_{k,l}}{\hat{S}^{2\gamma}_{k,l}} \right|^{\alpha}, \tag{3}$$

where $\hat{S}_{k,l}$ denotes the cepstrally smoothed version of $|S_{k,l}|$ and α and γ are hyper-parameters controlling the spiky nature of the resulting spectrum [18].

The IF spectrum is defined as the derivative of phase across time and can be computed as [32, 33]

$$\mathrm{IF}_{k,l} = \arg\{S_{k,l+1} S^*_{k,l}\}, \tag{4}$$

with $\arg\{\cdot\}$ denoting the complex phase function and $\{\cdot\}^*$ denoting the complex conjugate. Using (4) to compute the IF spectrum helps to partly alleviate phase wrapping issues [32, 33].

Figs. 1(c) and 1(d) depict the MGD and IF spectra for the previously considered exemplary utterance $s(n)$, with magnitude and phase spectra depicted in Figs. 1(a) and 1(b). It can be observed that contrary to the phase spectrum and similarly to the magnitude spectrum, both the MGD and IF spectra exhibit regular spectro-temporal patterns which can be potentially exploited by deep learning-based dysarthric speech detection approaches. Since the MGD and IF spectra result in magnitude-like representations of the phase spectrum, an experimental investigation is necessary to establish whether such representations provide complementary cues for dysarthric speech detection that cannot be extracted from the magnitude spectrum.

2.2. Temporal envelope and fine structure representations

Instead of computing signal representations based on the uniform STFT filter bank, in [31] we have proposed to compute the temporal envelope and fine structure representation using Gammatone filter

banks mimicking cochlear frequency analysis. To this end, the signal $s(n)$ is split into K complementary frequency bands of equal width along the human basiliar membrane [34]. Let us denote by $s_k^c(n)$ the signal obtained at the output of the k-th band pass filter. The analytic representation of $s_k^c(n)$ is given by

$$s_k^a(n) = s_k^c(n) + j\mathcal{H}\{s_k^c(n)\}, \tag{5}$$

where $\mathcal{H}\{\cdot\}$ denotes the Hilbert transform. The magnitude and cosine of the phase of the complex coefficients in (5) yield the temporal envelope and fine structure signals. These signals are sub-sampled by taking the mean over sliding windows of length N samples (with or without overlap) to obtain the final temporal envelope and fine structure representations used for dysarthric speech detection in [31].

3. MAGNITUDE AND PHASE-BASED DYSARTHRIC SPEECH DETECTION

To investigate the applicability of phase representations for dysarthric speech detection, we consider the state-of-the-art CNN-based approach from [10] depicted in Fig. 2(a). As shown in this figure, the CNN operates on $(K \times B)$–dimensional magnitude representations, with B denoting a user-defined number of time frames (cf. Section 4.4). Through alternating convolutional and max-pooling layers, the network learns a discriminative representation from the magnitude spectrum of neurotypical and dysarthric signals. In Section 4.4 we investigate the performance of this approach when $(K \times B)$–dimensional phase representations (i.e., unprocessed phase, MGD, IF) are used as input instead of the magnitude spectrum used in [10].

To further analyze whether phase representations provide additional cues that cannot be extracted from the magnitude, we consider the dual input CNN from [31] depicted in Fig. 2(b). In [31], this dual input CNN operates on $(K \times B)$–dimensional envelope and fine structure representations computed as described in Section 2.2. As shown in this figure, different convolutional and max-pooling layers are used on the different input representations. Hence, two different discriminative representations are learned and jointly exploited through fully-connected layers to detect dysarthric speech. Instead of using the temporal envelope and fine structure representations, in this paper we investigate the performance of the dual input CNN operating on the STFT magnitude spectrum and phase representations, i.e., magnitude and phase spectra, magnitude and MGD spectra, or magnitude and IF spectra.

4. EXPERIMENTAL RESULTS

In this section, the dysarthric speech detection performance of the single and dual input CNNs is compared for different input representations.

4.1. Database

We consider recordings of 24 different words and a phonetically balanced text from 50 neurotypical speakers and 50 PD patients from the well-balanced PC-GITA database [35]. The recordings are downsampled to 16 kHz from the original sampling frequency of 44.1 kHz. The average length of the total speech material available for each speaker is 32.1 seconds.

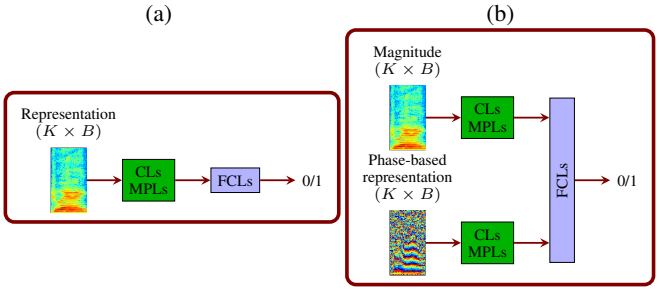

Fig. 2. Block diagram of the considered CNN-based dysarthric speech detection approach: (a) single input approach operating on the magnitude, phase, MGD, or IF spectrum and (b) dual input approach operating on the magnitude and phase spectra, magnitude and MGD spectra, or magnitude and IF spectra. CLs, MPLs, and FCLs refer to convolutional, max-pooling, and fully-connected layers, respectively.

4.2. Input representations and network architectures

The STFT is obtained using a weighted overlap-add framework with a Hanning analysis window without overlap and a frame size of $N = 160$ samples (i.e., 10 ms), resulting in $K = 81$ subbands. The logarithm of the magnitude spectrum, the phase spectrum, and the IF spectrum are straightforwardly computed from the STFT coefficients (cf. Section 2.1). To compute the MGD spectrum, we use a cepstral window of length 20 samples, $\alpha = 0.6$, and $\gamma = 0.3$ (cf. (3)). The temporal envelope and fine structure representations are also computed using 10 ms segments without overlap and $K = 81$ auditory filter banks. The remainder of the parameters used in computing the temporal envelope and fine structure representations are the same as in [31]. Similarly to [31], $(K \times B)$–dimensional segments using $B = 50$ and a 50% overlap are extracted from the computed representations and used as inputs to the single and dual input CNNs. Input representations are normalized to a mean of 0 and a standard deviation of 1.

The architecture of the single input CNN in Fig. 2(a) consists of two convolutional layers with 64 channels, a 2×2 kernel for the first layer, and a 3×3 kernel for the second layer. Each convolutional layer is followed by a ReLU activation function, batch normalization, and max-pooling with a 2×2 kernel. The second convolutional layer is followed by a dropout layer with a rate of 0.5. After the dropout layer, a fully-connected layer (input dimension of 4224 and output dimension of 2) followed by the softmax function is used.

The dual input CNN in Fig. 2(b) has the same architecture of convolutional, max-pooling, and dropout layers for the upper and lower branches as the single input CNN. The output of these two branches is fused through a fully-connected layer with an input size of 8448, an output size of 128, and a ReLU activation function. A final fully-connected layer with an input size of 128 and an output size of 2 followed by the softmax function is then used.

4.3. Training and evaluation

The performance of the considered approaches is evaluated in a speaker-independent stratified 10-fold cross-validation framework. The stochastic gradient descent algorithm and cross-entropy loss are used for training. The batch size is 128 and the initial learning rate is 0.01. In each training fold, a development set with the same size as the test set is used such that the learning rate is halved if the loss on

Table 1. Performance of the single input CNN operating on magnitude and phase representations.

Representation	Accuracy	AUC
Magnitude	69.72 ± 15.62	0.77 ± 0.16
Phase	62.76 ± 14.52	0.70 ± 0.15
MGD	70.78 ± 12.22	$\mathbf{0.79 \pm 0.12}$
IF	$\mathbf{72.64 \pm 13.37}$	$\mathbf{0.79 \pm 0.13}$

the development set does not decrease for 5 consecutive iterations. Training is stopped when the learning rate has decreased beyond 10^{-6} or after 100 epochs. The single input CNNs are randomly initialized. The convolutional layers of the dual input CNNs are initialized with the convolutional layers of the trained single input counterparts.

The final prediction score for a test speaker is obtained through soft voting of the prediction scores obtained for each $(K \times B)$–dimensional input representation belonging to the speaker. 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 that the random initialization of networks and the random split of speakers into training and testing folds have on the final performance, we have trained all networks with 5 different random seeds for 5 different splits of speakers. The reported performance measures are the mean and standard deviation of the performance obtained across these different models.

4.4. Results

To investigate the applicability of phase representations in comparison to the traditionally used magnitude spectrum, we analyze the performance of the single input CNN operating on different input representations. Table 1 presents the performance of the single input CNN operating on the magnitude, phase, MGD, and IF spectra. It can be observed that using the IF spectrum yields the highest performance in terms of accuracy, with an AUC score similar to the AUC score obtained when using the MGD spectrum. Further, it can be observed that the performance when using the magnitude and MGD spectra is very similar. This result is to be expected since as it can be visually inspected in Fig. 1, the MGD spectrum contains spectro-temporal patterns that are similar to the magnitude spectrum. The lowest performance in terms of accuracy and AUC score is obtained when using the phase spectrum, which is also to be expected since the phase spectrum is irregular and visually void of meaningful structures. However, it should be noted that using the phase spectrum yields an accuracy of 62.76% and an AUC score of 0.70, which shows that although the phase spectrum does not visually exhibit any regular spectro-temporal structures, a CNN nevertheless manages to partially discover cues in the phase spectrum that are important for dysarthric speech detection.

To investigate whether phase representations provide complementary cues that cannot be extracted from the magnitude, we analyze the performance of the dual input CNN operating on the magnitude and different phase representations. Table 2 presents the performance of the dual input CNN operating on the magnitude and phase spectra, the magnitude and MGD spectra, and the magnitude and IF spectra. In addition, the performance of the dual input CNN from [31] operating on the temporal envelope and fine structure representations is also presented. When comparing the dual input CNNs operating on different magnitude and phase representations, it can be observed that using the magnitude and IF spectra yields the highest performance, with an accuracy of 93.69% and an AUC score of 0.97.[1] Further, it can be observed that combining the magnitude and phase spectra yields a considerably better performance than combining the magnitude and MGD spectra. This result shows that although the phase spectrum is irregular, it contains more complementary cues to the magnitude spectrum for dysarthric speech detection than the MGD spectrum. A comparison of the results in Tables 1 and 2 shows that all phase representations contain complementary cues to the magnitude spectrum, with all dual input CNNs yielding a considerably better performance than their single input counterparts. Finally, Table 2 shows that using the temporal envelope and fine structure representations yields a similar performance as using the magnitude and unprocessed phase representations, but a considerably worse performance than using the magnitude and IF representations. These results confirm that the performance improvement we obtained in [31] can be attributed to the incorporation of the analytical phase of the signal and not to the use of auditory-inspired filter banks. Nevertheless, exploring alternative representations of the temporal fine structure signals that are applicable for dysarthric speech detection remains a viable future research direction.

Table 2. Performance of the dual input CNN operating on the magnitude spectrum and different phase representations. The performance of the dual input CNN from [31] operating on the temporal envelope and fine structure signals is also presented.

Representation	Accuracy	AUC
Magnitude-Phase	87.32 ± 9.69	0.93 ± 0.10
Magnitude-MGD	80.92 ± 10.11	0.90 ± 0.10
Magnitude-IF	$\mathbf{93.68 \pm 5.32}$	$\mathbf{0.97 \pm 0.05}$
Envelope-Fine structure	86.04 ± 8.03	0.94 ± 0.08

In summary, the results presented in this section confirm that all phase representations of the STFT provide useful cues for dysarthric speech detection and should be used in addition to the traditionally used magnitude representation. In particular, combining the magnitude and IF spectra results in a very high dysarthric speech detection accuracy.

5. CONCLUSION

Deep learning-based dysarthric speech detection approaches typically learn discriminative representations by processing the magnitude spectrum of signals and ignoring the phase spectrum. In this paper we have investigated the applicability of STFT phase representations for dysarthric speech detection. Since the phase spectrum is irregular and visually void of spectro-temporal patterns, we have analyzed two alternative representations which reveal hidden structures of the phase spectrum, i.e., the MGD and IF spectra. Using a single input CNN we have shown that all considered phase representations, i.e., the unprocessed phase, MGD, and IF spectra, contain dysarthric cues. Using a dual input CNN operating on both the magnitude and phase representations we have shown that all considered phase representations serve as complementary features to the magnitude spectrum, with the combination of magnitude and IF spectra yielding a very high performance. The presented results have demonstrated the importance of considering phase information for dysarthric speech detection and will hopefully motivate research on novel architectures to optimally combine the magnitude and phase information.

[1] It should be noted that to the best of our knowledge, this is the highest performance that has been reported using deep learning-based dysarthric speech detection approaches on this speech material from the PC-GITA database.

6. REFERENCES

[1] J. R. Duffy, *Motor speech disorders: substrates, differential diagnosis, and management*, Elsevier Mosby, Missouri, USA, 2003.

[2] J. Rusz, R. Cmejla, T. Tykalova, H. Ruzickova, J. Klempir, V. Majerova, J. Picmausova, J. Roth, and E. Ruzicka, "Imprecise vowel articulation as a potential early marker of Parkinson's disease: Effect of speaking task," *Journal of the Acoustical Society of America*, vol. 134, no. 3, pp. 2171–2181, Sept. 2013.

[3] J. M. Tracy, Y. Özkanca, D. C. Atkins, and R. H. Ghomi, "Investigating voice as a biomarker: Deep phenotyping methods for early detection of Parkinson's disease," *Journal of Biomedical Informatics*, vol. 104, Apr. 2020.

[4] 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.

[5] 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, Sept. 2015, pp. 95–99.

[6] I. Kodrasi and H. Bourlard, "Spectro-temporal sparsity characterization for dysarthric speech detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 28, no. 1, pp. 1210–1222, Apr. 2020.

[7] P. Janbakhshi, I. Kodrasi, and H. Bourlard, "Subspace-based learning for automatic dysarthric speech detection," *IEEE Signal Processing Letters*, vol. 28, pp. 96–100, Dec. 2020.

[8] 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, Sept. 2020, pp. 2897–2901.

[9] 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.

[10] 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.

[11] 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.

[12] 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, Sept. 2018, pp. 1913–1917.

[13] J. C. Vasquez-Correa, T. Arias-Vergara, M. Schuster, J. R. 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, Sept. 2020.

[14] P. Janbakhshi and I. Kodrasi, "Supervised speech representation learning for Parkinson's disease classification," in *Proc. ITG conference on Speech Communication*, Kiel, Germany, Sept. 2021, pp. 154–158.

[15] P. Mowlaee, R. Saeidi, and Y. Stylianou, "INTERSPEECH 2014 Special Session: Phase importance in speech processing applications," in *Proc. Annual Conference of the International Speech Communication Association*, Singapore, Sept. 2014, pp. 1623–1627.

[16] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, "Phase processing for single-channel speech enhancement: History and recent advances," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 55–66, Mar. 2015.

[17] B. Yegnanarayana, H. A. Murthy, and V. R. Ramachandran, "Speech enhancement using group delay functions," in *Proc. International Conference on Spoken Language Processing*, Kobe, Japan, Nov. 1990, pp. 301–304.

[18] H. A. Murthy and V. Gadde, "The modified group delay function and its application to phoneme recognition," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Hong Kong, Apr. 2003, pp. 68–71.

[19] D. Friedman, "Instantaneous-frequency distribution vs. time: An interpretation of the phase structure of speech," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Florida, USA, Mar. 1985, pp. 1121–1124.

[20] B. Boashash, "Estimating and interpreting the instantaneous frequency of a signal – Part 1: Fundamentals," *Proceedings of the IEEE*, vol. 80, no. 4, pp. 520–538, May 1992.

[21] A. V. Oppenheim J. S. Lim, "The importance of phase in signals," *Proceedings of the IEEE*, vol. 69, no. 5, pp. 529–541, May 1981.

[22] D. Wang and J. Lim, "The unimportance of phase in speech enhancement," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, no. 4, pp. 679–681, Aug. 1982.

[23] T. Gerkmann and M. Krawczyk, "MMSE-optimal spectral amplitude estimation given the STFT-phase," *IEEE Signal Processing Letters*, vol. 20, no. 2, pp. 129–132, Feb. 2013.

[24] R. M. Hegde, H. A. Murthy, and V. R. R. Gadde, "Significance of the modified group delay feature in speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 190–202, Jan. 2007.

[25] S. Sehgal, S. Cunningham, and P. Green, "Phase-based feature representations for improving recognition of dysarthric speech," in *Proc. IEEE Spoken Language Technology Workshop*, Athens, Greece, Dec. 2018, pp. 13–20.

[26] G. Degottex, A. Roebel, and X. Rodet, "Phase minimization for glottal model estimation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1080–1090, July 2011.

[27] S. Nakagawa, L. Wang, and S. Ohtsuka, "Speaker identification and verification by combining MFCC and phase information," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 4, pp. 1085–1095, May 2012.

[28] P. Rajan, T. Kinnunen, C. Hanilci, J. Pohjalainen, and P. Alku, "Using group delay functions from all-pole models for speaker recognition," in *Proc. Annual Conference of the International Speech Communication Association*, Lyon, France, Aug. 2013, pp. 2489–2493.

[29] J. Deng, X. Xu, Z. Zhang, S. Fruehholz, and B. Schuller, "Exploitation of phase-based features for whispered speech emotion recognition," *IEEE Access*, vol. 4, pp. 4299–4309, July 2016.

[30] L. Guo, L. Wang, J. Dang, L. Zhang, H. Guan, and X. Li, "Speech emotion recognition by combining amplitude and phase information using convolutional neural network," in *Proc. Annual Conference of the International Speech Communication Association*, Hyderabad, India, Sept. 2018, pp. 1611–1615.

[31] I. Kodrasi, "Temporal envelope and fine structure cues for dysarthric speech detection using CNNs," *IEEE Signal Processing Letters*, vol. 28, pp. 1853–1857, Aug. 2021.

[32] S. Kay, "A fast and accurate single frequency estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 12, pp. 1987–1990, Dec. 1989.

[33] A. Stark and K. Paliwal, "Speech analysis using instantaneous frequency deviation," in *Proc. Annual Conference of the International Speech Communication Association*, Brisbane, Australia, Jan. 2008, pp. 2602–2605.

[34] 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.

[35] 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.