Impact of Speech Mode in Automatic Pathological Speech Detection

Shakeel A. Sheikh, Ina Kodrasi
Signal Processing for Communication Group, Idiap Research Institute, Martigny, Switzerland
{shakeel.sheikh, ina.kodrasi}@idiap.ch

***Abstract*—Automatic pathological speech detection approaches yield promising results in identifying various pathologies. These approaches are typically designed and evaluated for phonetically-controlled speech scenarios, where speakers are prompted to articulate identical phonetic content. While gathering controlled speech recordings can be laborious, spontaneous speech can be conveniently acquired as potential patients navigate their daily routines. Further, spontaneous speech can be valuable in detecting subtle and abstract cues of pathological speech. Nonetheless, the efficacy of automatic pathological speech detection for spontaneous speech remains unexplored. This paper analyzes the influence of speech mode on pathological speech detection approaches, examining two distinct categories of approaches, i.e., classical machine learning and deep learning. Results indicate that classical approaches may struggle to capture pathology-discriminant cues in spontaneous speech. In contrast, deep learning approaches demonstrate superior performance, managing to extract additional cues that were previously inaccessible in non-spontaneous speech.**

***Index Terms*—pathological speech, spontaneous speech, non-spontaneous speech, deep learning.**

I. Introduction

Pathological speech resulting from neurological disorders poses a significant healthcare challenge, where early diagnosis is crucial for delaying progression and for managing the condition. Pathological speech conditions such as dysarthria or apraxia of speech are characterized by alterations in speech production, including imprecise articulation, abnormal pitch, or irregular rhythm, making it challenging for patients to effectively convey their intentions through speech [1]–[4]. Automatic speech-based diagnostic approaches emerge as cost-effective methods for early detection [1], especially compared to traditional diagnostic techniques such as Positron emission tomography and Dopamine transporter scans, which are expensive and inaccessible, creating a barrier to timely diagnosis and treatment [5], [6]. Such approaches provide convenient monitoring of disease progression, reducing the need for costly in-person therapy sessions. As a result, various automatic approaches have been proposed for pathological speech detection, broadly falling into two categories: i) those utilizing classical machine learning techniques with hand-crafted acoustic features [7], [8], and ii) those based on deep learning architectures operating on time-frequency input representations [9]–[12] or self-supervised embeddings [13]–[16]. The first category of approaches often employs support vector machines (SVMs) with Mel-frequency cepstral coefficients (MFCCs) [17], openSMILE features [14], or sparsity-based features [18]. In contrast, the second category of approaches focuses on exploring time-frequency input representations with different network architectures and training paradigms such as long short-term memory networks [19], autoencoders (AEs) [20], adversarial training [9], or convolutional neural networks (CNNs) [9], [21]. More recently, approaches in the second category focus on using self-supervised embeddings such as wav2vec2 for automatic pathological speech detection [13], [14].

While both categories of automatic approaches have demonstrated promising results, they are typically developed and assessed in phonetically-controlled, non-spontaneous, speech settings. Phonetically-controlled non-spontaneous speech refers to elicited speech, where speakers are prompted to repeat the same words, phrases, or sentences, meticulously designed by clinicians to elicit discernible cues indicative of pathological speech. In contrast, spontaneous speech reflects natural conversation and can be valuable in detecting subtle and abstract cues of pathological speech. Furthermore, spontaneous speech mirrors real-world communication and places a higher cognitive load on real-time processing of various speech tasks such as planning, precise sensorimotor execution, and articulation, making it more susceptible to deficits related to pathology [1]. Yet, to the best of our knowledge, the performance of state-of-the-art automatic pathological speech detection approaches on spontaneous speech remains unexplored. Further investigation is required to assess how these approaches fare in more natural settings, where speech is less structured and influenced by diverse factors such as individual variations in communication styles.

This paper analyzes the effect of *speech mode*, i.e., non-spontaneous and spontaneous speech, on the performance of numerous state-of-the-art automatic pathological speech detection approaches. The evaluation is conducted on two distinct databases, i.e., the Spanish PC-GITA database containing dysarthria recordings from patients with Parkinson's disease [22] and the French MoSpeeDi database [23] containing dysarthria recordings from patients with Parkinson's disease or Amytrophic Lateral Sclerosis. The considered classical machine learning-based approaches include SVMs operating on hand-crafted acoustic features like openSMILE [24], MFCCs [17], and sparsity-based features [18]. Additionally, the considered deep learning-based approaches include the CNN-based approach proposed in [21], the AE-based approach proposed in [9], and the wav2vec2-based approach in [14].

II. METHODS

This section outlines the state-of-the-art pathological speech detection approaches considered in this paper for analyzing the impact of the speech mode on performance.

A. Input Representation

In the following, we describe the different input representations used in the considered approaches.

OpenSMILE: OpenSMILE features are commonly used as input to classical machine learning approaches [14], [25]. For each utterance, we extract a 6373–dimensional feature vector using the openSMILE toolkit [24]. Following feature extraction, we employ principal component analysis for dimensionality reduction similar to previous works [14], [25]. We retain only the features that explain 95% of the variance in the training data.

MFCCs: MFCC features are also commonly used as input to classical machine learning approaches [8]. For each utterance, we compute the mean, variance, kurtosis, and skewness of the first 12 MFCC coefficients, resulting in a 48–dimensional feature vector as in [14]. The MFCC features are extracted using the openSMILE toolkit [24].

Sparsity-based features: Sparsity-based features have also been used with classical machine learning approaches [18], [26]. Following the methodology outlined in [18], [26], we compute sparsity-based features for each utterance using the shape parameter of a Chi distribution. The process involves computing the short-time Fourier transform (STFT) using a Hamming window of length 32 ms and a hop size of 4 ms. Next, we obtain the maximum likelihood estimate of the shape parameter for the Chi distribution that best models the spectral magnitude at each frequency bin. This procedure results in a 257-dimensional feature vector for each utterance.

Mel spectrograms: Mel spectrograms are typically used with deep learning-based approaches [8]. For each speech segment, we compute the STFT coefficients using a Hamming window of length 32 ms and a hop size of 4 ms. The computed coefficients are converted into Mel-scale representations with 126 Mel-bands, following the approach used in [14]. The logarithm of the Mel spectrogram is then used as the input representation for deep learning-based approaches.

Self-supervised embeddings: The wav2vec2 framework has revolutionised the extraction of meaningful contextual representations from raw speech [27]. These extracted embeddings have shown promising performance in various pathological speech tasks [13], [28]–[30]. For our analysis, we use embeddings extracted from the XLSR-53 model [31]. It has been shown that the last layers of wav2vec2 models are more adapted towards non paralinguistic and other phonetic content-related tasks, while the first layers are more applicable to paralinguistic and prosody-related tasks [27], [30], [32]. Hence, in this paper, we consider embeddings extracted from the first transformer layer of the XLSR-53 model. For each utterance, the final input representation is computed as the mean and standard deviation of the embeddings across time. This procedure results in a 2048–dimensional input representation. It should be noted that the wav2vec2 model is not fine-tuned, but rather utilized as a self-supervised feature extractor for automatic pathological speech detection.

B. Classifiers

In the following, we describe the different classifiers used in the considered approaches.

Support vector machines: To analyze the performance of classical approaches based on handcrafted acoustic features, we use SVMs with radial basis kernel function as in [14], [17], [18]. Different SVMs are trained for different acoustic features, i.e., for openSMILE, MFCCs, and sparsity-based features.

Convolutional neural networks: CNNs have been widely used in various speech applications, including automatic pathological speech detection [12], [33]–[35]. For our analysis, we train a CNN with Mel spectrogram input representations as in [14]. The CNN consists of a normalization layer, followed by two convolutional layers with 64 channels each and kernel sizes of 2×2 and 3×3 respectively. Each convolutional layer is followed by batch normalisation, max-pooling with a 2×2 or 3×3 kernel, and a ReLU activation function. A dropout rate of 30% is applied after the second convolutional layer. Finally, the output is fed to a fully-connected linear layer (input size: 25600, output size: 2) for pathological speech detection.

Autoencoders: Given the promising performance of the AE-based framework for pathological speech detection in [9], we also consider this approach in our analysis. Also in this framework Mel spectrograms are used as the input representation. The encoder θ_e is composed of 4 convolutional layers with a kernel size of 3×3. Each convolutional layer is followed by batch normalization, max-pooling with a 2×2 kernel, and a ReLU activation function. The output of the encoder is fed to a bottleneck layer of size 128. This bottleneck representation is then decoded by the decoder θ_d to reconstruct the input representation. The decoder mirrors the encoder components in reverse order, using interpolation and transposed convolution instead of max-pooling and convolution. In addition, the bottleneck representation is simultaneously fed to a classifier aiming to learn a pathology-discriminant representation in a multi-task learning fashion. The pathology classifier θ_{pa} is composed of a fully connected linear layer with 128 input units and 2 output units. The classification loss $\mathcal{L}_{pa}$ and the AE reconstruction loss $\mathcal{L}_{ae}$ are jointly minimized by optimizing the parameters θ_e, θ_d, θ_{pa} using

$$\mathcal{L}(\hat{\theta}_e, \hat{\theta}_d, \hat{\theta}_{pd}) = \underset{\theta_e, \theta_d, \theta_{pd}}{\operatorname{argmin}} \; \mathcal{L}(\theta_e, \theta_d, \theta_{pd}), \tag{1}$$

where

$$\mathcal{L}(\theta_e, \theta_d, \theta_{pd}) = (1-\alpha)\mathcal{L}_{ae}(\theta_e, \theta_d) + \alpha \mathcal{L}_{pd}(\theta_e, \theta_{pd}), \tag{2}$$

with $\alpha \in [0, 1]$ being a trade-off parameter between the AE loss and the classification loss. In our analysis, we use $\alpha = 0.01$ as in [14]. For inference, the decoder is disregarded and the bottleneck representation is directly fed to the pathology classifier.

Fully connected linear layer: Similar to [14], we employ a fully connected linear layer for pathological speech classification when using wav2vec2 embeddings as the input representation. The linear layer consists of 2048 input units and 2 output units.

III. Experimental Settings

In this section, the settings used for the experimental analysis are provided.

A. Databases

Analysis are conducted using two different databases, i.e., the Spanish PC-GITA database [22] and the French MoSpeeDi database [23].

PC-GITA: The PC-GITA database contains Spanish recordings from a gender-balanced group of 50 patients diagnosed with Parkinson's disease along with a gender-balanced group of 50 neurotypical speakers. For the non-spontaneous speech mode, we use recordings of 10 sentences and a phonetically balanced text. Using these recordings, the average length of the total available non-spontaneous speech material across speakers is 55.4 s. For the spontaneous speech mode, we use recordings of the speakers talking about what they do in a normal day. Using these recordings, the average length of the available spontaneous speech material across speakers is 47.1 s.

MoSpeeDi: From the French MoSpeeDi database, we use recordings from a gender-balanced group of 35 patients diagnosed with Parkinson's disease or Amytrophic Lateral Sclerosis along with a gender-balanced group of 35 neurotypical speakers. For the non-spontaneous speech mode, we use recordings of 8 sentences. Using these recordings, the average length of the total available non-spontaneous speech material across speakers is 97.7 s. For the spontaneous speech mode, we use recordings of the speakers talking about their holidays. Using these recordings, the average length of the available spontaneous speech material across speakers is 153.1 s.

B. Evaluation and Training

For all considered approaches, evaluation is done within a stratified K-fold cross validation strategy, with $K = 10$ for the PC-GITA database and $K = 7$ for the MoSpeeDi database. For each fold of the PC-GITA database, we use 80%, 10%, and 10% of the data for training, validation, and testing, respectively. For each fold of the MoSpeeDi database, we use 72%, 14%, and 14% of the data for training, validation, and testing, respectively.

The SVM-based and the wav2vec2-based approaches accept variable length segments of speech as input, hence, full utterances are used as input to these approaches. The CNN-based and the AE-based approach accept only fixed-size segments of speech as input. For these two approaches, we segment available utterances into 500 ms segments with a 50% overlap and use these fixed-size segments as input.

The performance is evaluated in terms of speaker-level accuracy. For the SVM-based approaches, speaker-level accuracy is computed through majority voting of the decisions for all utterances belonging to each speaker. For the remaining approaches, speaker-level accuracy is computed through soft voting of the probability of decisions for all segments/utterances belonging to each speaker.

Implementation is done using PyTorch and Torchaudio. For the SVM-based approaches, separate SVMs are trained for each of the handcrafted acoustic features, i.e., for openSMILE, MFCCs, and sparsity-based features. By optimizing the performance on the validation set, the optimal kernel width γ and soft margin constant $\mathcal{C}$ is found from the sets $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}]$ and $\{10, 10^2, 10^3, 10^4\}$ respectively. The remaining approaches are trained using the Adam optimizer. The initial learning rates are 0.0015 for the AE-based approach, 0.001 for the CNN-based approach, and 0.001 for the wav2vec2-based approach. The learning rates are reduced after every 15 epochs using the *MultiStepLR* scheduler with γ = 0.9. Training is terminated if the validation loss does not decrease for 10 consecutive epochs, with the last best model saved and used for testing.

IV. Experimental Results

In this section, we analyse and discuss the impact of speech modes (i.e., non-spontaneous and spontaneous) in SVM-based approaches, CNN- and AE-based approaches, and the wav2vec2-based approach. Results obtained for all considered approaches and both databases are presented in Table I.

SVM-based approaches

As shown in Table I, the different SVM-based approaches do not consistently outperform in one speech mode over the other. On the one hand, using openSMILE features yields a better performance on spontaneous speech compared to non-spontaneous speech, regardless of the database. However, openSMILE features are designed for various tasks like speech emotion recognition or audio detection, which is not ideal for pathological speech detection. Consequently, many openSMILE features capture cues unrelated to pathology. This is reflected in the overall low performance achieved by openSMILE features, regardless of the speech mode or database used. On the other hand, MFCC features demonstrate better performance on spontaneous speech than on non-spontaneous speech in the MoSpeeDi database, while the opposite trend is observed in the PC-GITA database. In spontaneous speech, the various acoustic, linguistic, spectral, and phonetic features exhibit a lot more non-pathology related fluctuations than in non-spontaneous settings. It is expected that MFCC features inadvertently capture phonetic variations in spontaneous recordings, resulting in lower performance for spontaneous speech than for non-spontaneous speech in the PC-GITA database. The absence of this performance trend in the MoSpeeDi database warrants further investigation. Finally, using sparsity-based features yields a better performance on non-spontaneous speech than on spontaneous speech, regardless of the database. This result is expected, given that sparsity-based features are derived by fitting a distribution to the speech

TABLE I
SPEAKER-LEVEL CLASSIFICATION ACCURACY [%] OF THE CONSIDERED AUTOMATIC PATHOLOGICAL SPEECH DETECTION APPROACHES ON THE PC-GITA AND MOSPEEDI DATABASE USING NON-SPONTANEOUS AND SPONTANEOUS SPEECH RECORDINGS.

Approach	PC-GITA		MoSpeeDi	
	Non-spotaneous	Spontaneous	Non-spotaneous	Spontaneous
SVM with openSMILE input features	43.0 ± 17.0	57.0 ± 17.0	42.9 ± 7.76	57.1 ± 19.8
SVM with MFCC input features	68.0 ± 09.2	62.0 ± 13.2	71.4 ± 10.7	84.3 ± 15.1
SVM with sparsity-based input features	73.0 ± 09.5	56.0 ± 22.2	82.8 ± 11.1	78.6 ± 09.0
CNN with Mel spectrogram input representations	75.0 ± 07.0	74.0 ± 12.6	90.1 ± 11.5	92.9 ± 11.1
AE with Mel spectrogram input representations	71.0 ± 13.7	80.0 ± 14.1	82.9 ± 16.0	85.7 ± 05.3
Linear layer with wav2vec2 embedding input representations	77.0 ± 14.9	75.0 ± 15.8	87.1 ± 11.12	87.1 ± 11.1

spectral coefficients. The use of single utterances for this fitting process influences the quality of the fit, leading to distributions that are greatly influenced by the phonetic content of the utterance. The fluctuating phonetic content among spontaneous speech recordings results in unreliable distribution fits that not only exhibit pathological cues, but also phonetic content cues.

In summary, results show that the performance of different SVM-based approaches is differently influenced by the speech mode, depending on the used handcrafted acoustic features. These findings emphasize the importance of carefully crafting and selecting acoustic features for classical approaches, such that undesirable cues are not captured.

CNN-based and AE-based approaches

While SVM-based approaches are not powerful enough to ignore pathology-unrelated fluctuations in spontaneous speech, the CNN-based and AE-based approaches have a better potential. Results in Table I show that the CNN-based approach yields a similar performance for both speech modes in both databases whereas the AE-based approach performs better on spontaneous speech recordings than on non-spontaneous ones. These findings indicate that these approaches can not only disregard pathology-unrelated fluctuations in spontaneous speech, but can also extract additional cues that were not identifiable in non-spontaneous speech. It should be noted that unlike the SVM-based approaches operating on a single utterance, the CNN- and AE-based approaches operate on multiple segments of speech, even for the spontaneous speech recordings. The availability of more samples when using the spontaneous speech recordings for the CNN- and AE-based approaches might contribute to the observed improved performance.

wav2vec2-based approach

Similar to the CNN- and AE-based approaches, we anticipate the wav2vec2-based approach to be able to disregard pathology-unrelated variations in spontaneous speech. As expected, the results in Table I demonstrate that the wav2vec2-based approach performs similarly for both speech modes and databases, with a negligible decrease in performance for spontaneous speech in the PC-GITA database. The XLRS53 variant of the wav2vec2 model is trained on a diverse range of non-spontaneous and spontaneous multilingual databases [31], resulting in embeddings that do not exhibit significant fluctuations caused by the speech mode.

Overall, the results presented in this section show that on the one hand, SVM-based approaches struggle to perform well in spontaneous speech recordings depending on the used acoustic features. On the other hand, deep learning-based approaches not only have the ability to disregard pathology-unrelated fluctuations in spontaneous speech, but can also extract additional pathology-discriminant cues that may not be present in non-spontaneous speech. In the future, we will investigate how the performance of the SVM-based and wav2vec2-based approaches changes when the spontaneous speech recording is segmented into shorter fixed-sized segments (as for the CNN- and AE-based approaches), rather than treating it as a single utterance. Furthermore, we will investigate the extraction of phonetic-dependent acoustic features from spontaneous speech to be used in classical approaches.

V. CONCLUSION

Pathological speech detection approaches developed so far have been trained and evaluated only in phonetically-controlled non-spontaneous speech settings. However, spontaneous speech can be more conveniently recorded and it can be valuable in detecting subtle and abstract cues of pathological speech. In this paper, we have examined the performance of various classical and deep learning-based pathological speech detection approaches using both non-spontaneous and spontaneous speech recordings. Our findings suggest that classical approaches employing handcrafted acoustic features may encounter challenges in capturing pathology-discriminant cues in spontaneous speech. Conversely, deep learning approaches exhibit superior performance, successfully extracting additional cues that were previously inaccessible in non-spontaneous speech. Future work will target further improving the performance of different approaches on spontaneous speech recordings.

VI. ACKNOWLEDGEMENTS

This work was supported by the Swiss National Science Foundation project CRSII5_202228 on “Characterisation of motor speech disorders and processes”.

REFERENCES

[1] J. S. Damico, N. Müller, and M. J. Ball, *The Handbook of Language and Speech Disorders*. Wiley Online Library, 2010.

[2] C. Stewart, L. Winfield, A. Hunt, S. B. Bressman, S. Fahn, A. Blitzer, and M. F. Brin, "Speech dysfunction in early Parkinson's disease," *Movement Disorders: Journal of the Movement Disorder Society*, vol. 10, no. 5, pp. 562–565, Sept. 1995.

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

[4] L. Baghai-Ravary and S. W. Beet, *Automatic speech signal analysis for clinical diagnosis and assessment of speech disorders*. Springer Science & Business Media, Aug. 2012.

[5] D. J. Brooks, "Detection of preclinical Parkinson's disease with PET," *Geriatrics*, vol. 46, no. 1, pp. 25–30, Aug. 1991.

[6] F. J. Martínez-Murcia, J. M. Górriz, J. Ramírez, I. Illán, A. Ortiz, Parkinson's Progression Markers Initiative *et al.*, "Automatic detection of Parkinsonism using significance measures and component analysis in datscan imaging," *Neurocomputing*, vol. 126, pp. 58–70, Feb 2014.

[7] N. Narendra and P. Alku, "Dysarthric speech classification using glottal features computed from non-words, words and sentences." in *Proc. Annual Conference of the International Speech Communication*, Hyderabad, India, Sept. 2018, pp. 3403–3407.

[8] A. A. Joshy and R. Rajan, "Automated dysarthria severity classification: A study on acoustic features and deep learning techniques," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 1147–1157, May 2022.

[9] 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. 1–5.

[10] J. Vasquez-Correa, T. Arias-Vergara, M. Schuster, J. Orozco-Arroyave, and E. Nöth, "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.

[11] P. Janbakhshi and I. Kodrasi, "Experimental investigation on stft phase representations for deep learning-based dysarthric speech detection," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Singapore, May 2022, pp. 6477–6481.

[12] T. Bhattacharjee, J. Mallela, Y. Belur, A. Nalini, R. Yadav, P. Reddy, D. Gope, and P. K. Ghosh, "Source and vocal tract cues for speech-based classification of patients with Parkinson's disease and healthy subjects." in *Proc. Annual Conference of the International Speech Communication*, Brnu, Czech Republic, Sept. 2021, pp. 2961–2965.

[13] F. Javanmardi, S. Tirronen, M. Kodali, S. R. Kadiri, and P. Alku, "Wav2vec-based detection and severity level classification of dysarthria from speech," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Rhodes Island, Greece, June 2023, pp. 1–5.

[14] G. Schu, P. Janbakhshi, and I. Kodrasi, "On using the UA-speech and Torgo databases to validate automatic dysarthric speech classification approaches," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Rhodes Island, Greece, June 2023, pp. 1–5.

[15] S. A. S, M. Sahidullah, F. Hirsch, and S. Ouni, "Machine learning for stuttering identification: Review, challenges and future directions," *Neurocomputing*, 2022.

[16] F. Javanmardi, S. Tirronen, M. Kodali, S. R. Kadiri, and P. Alku, "Wav2vec-based detection and severity level classification of dysarthria from speech," in *Proc. Annual Conference of the International Speech Communication*, Dublin, Ireland, Aug. 2023, pp. 1–5.

[17] J. R. Orozco-Arroyave, F. Hönig, J. D. Arias-Londoño, J. F. Vargas-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*, Dresden, Germany, Sept. 2015.

[18] I. Kodrasi and H. Bourlard, "Spectro-temporal sparsity characterization for dysarthric speech detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1210–1222, April 2020.

[19] J. Mallela, A. Illa, Y. Belur, A. Nalini, 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*, Shanghai, China, Oct. 2020, pp. 4586–4590.

[20] P. Janbakhshi and I. Kodrasi, "Adversarial-free speaker identity-invariant representation learning for automatic dysarthric speech classification," in *Proc. Annual Conference of the International Speech Communication*, Incheon, Korea, Sept. 2022, pp. 2138–2142.

[21] J. C. Vásquez-Correa, J. R. Orozco-Arroyave, and E. Nöth, "Convolutional neural network to model articulation impairments in patients with Parkinson's disease." in *Proc. Annual Conference of the International Speech Communication*, Stockholm, Sweden, Aug. 2017, pp. 314–318.

[22] J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, M. C. Gonzalez-Rátiva, and E. Nöth, "New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease." in *Proc. Language Resources and Evaluation Conefrence*, Reykjarik, Iceland, May 2014, pp. 342–347.

[23] "MoSpeeDi-ChaSpeePro dataset," https://www.unige.ch/fapse/mospeedi/mospeedi-dataset, database available within the MoSpeeDi-ChaSpeePro project consortium.

[24] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSmile: The Munich versatile and fast open-source audio feature extractor," in *Proc. ACM International Conference on Multimedia*, Feirenze, Italy, Oct. 2010, pp. 1459–1462.

[25] 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*, Shanghai, China, Oct. 2020, pp. 4991–4995.

[26] I. Kodrasi and H. Bourlard, "Statistical modeling of speech spectral coefficients in patients with Parkinson's disease," in *Proc. ITG Conference on Speech Communication*, Oldenburg, Germany, Sept. 2018, pp. 1–5.

[27] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Annual Conference on Neural Information Processing Systems*, vol. 33, Virtual Online, Dec. 2020, pp. 12 449–12 460.

[28] J. Švec, F. Polák, A. Bartoš, M. Zapletalová, and M. Víta, "Evaluation of wav2vec speech recognition for speakers with cognitive disorders," in *Proc. International Conference on Text, Speech, and Dialogue*, Brno, Czech Republic, Sept. 2022, pp. 501–512.

[29] Y. Getman, R. Al-Ghezi, E. Voskoboinik, T. Grósz, M. Kurimo, G. Salvi, T. Svendsen, and S. Strömbergsson, "Wav2vec2-based speech rating system for children with speech sound disorder," in *Proc. Annual Conference of the International Speech Communication*, Incheon, Korea, Sept. 2022, pp. 3618–3622.

[30] S. A. S, M. Sahidullah, S. Ouni, and F. Hirsch, "End-to-end and self-supervised learning for compare 2022 stuttering sub-challenge," in *Proc. ACM International Conference on Multimedia*, Lisbon, Portugal, Oct. 2022, pp. 7104–7108.

[31] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[32] J. Shah *et al.*, "What all do audio transformer models hear? Probing acoustic representations for language delivery and its structure," *arXiv preprint arXiv:2101.00387*, 2021.

[33] F. Javanmardi, S. R. Kadiri, and P. Alku, "Pre-trained models for detection and severity level classification of dysarthria from speech," *Speech Communication*, vol. 158, p. 103047, Feb. 2024.

[34] S. Gupta, A. T. Patil, M. Purohit, M. Parmar, M. Patel, H. A. Patil, and R. C. Guido, "Residual neural network precisely quantifies dysarthria severity-level based on short-duration speech segments," *Neural Networks*, vol. 139, pp. 105–117, July 2021.

[35] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, Jan. 2020.