ON USING THE UA-SPEECH AND TORGO DATABASES TO VALIDATE AUTOMATIC DYSARTHRIC SPEECH CLASSIFICATION APPROACHES

Guilherme Schu[⋆,†]*, Parvaneh Janbakhshi*[‡]*, Ina Kodrasi*[⋆]

[⋆]Idiap Research Institute, Martigny, Switzerland
[†]École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
[‡] Bayer AG, Berlin, Germany
guilherme.garcia@idiap.ch

ABSTRACT

Although the UA-Speech and TORGO databases of control and dysarthric speech are invaluable resources made available to the research community with the objective of developing robust automatic speech recognition systems, they have also been used to validate a considerable number of automatic dysarthric speech classification approaches. Such approaches typically rely on the underlying assumption that recordings from control and dysarthric speakers are collected in the same noiseless environment using the same recording setup. In this paper, we show that this assumption is violated for the UA-Speech and TORGO databases. Using voice activity detection to extract speech and non-speech segments, we show that the majority of state-of-the-art dysarthria classification approaches achieve the same or a considerably better performance when using the non-speech segments of these databases than when using the speech segments. These results demonstrate that such approaches trained and validated on the UA-Speech and TORGO databases are potentially learning characteristics of the recording environment or setup rather than dysarthric speech characteristics. We hope that these results raise awareness in the research community about the importance of the quality of recordings when developing and evaluating automatic dysarthria classification approaches.

***Index Terms*—** automatic dysarthria classification, TORGO, UA-Speech, noise, SNR

1. INTRODUCTION

Dysarthria is a motor speech disorder that occurs due to brain trauma or neurological conditions such as Cerebral Palsy (CP), Amyotrophic Lateral Sclerosis (ALS), or Parkinson's disease, and may affect the overall communicative ability of a patient [1]. To diagnose and manage it, speech pathologists perform auditory-perceptual assessments to evaluate different components of the speech production mechanism. However, these assessments can be time-consuming and subjective [2]. Aiming at assisting healthcare professionals, there has been a growing interest in the research community to develop automatic dysarthria classification approaches.

State-of-the-art automatic dysarthria classification approaches can be broadly grouped into two categories, i.e., i) approaches which use handcrafted features with classical machine learning classifiers [3–9] and ii) deep learning approaches which are trained to automatically extract and classify discriminative speech representations [10–17]. Commonly used approaches in the first category exploit support vector machines (SVMs) with Mel-frequency cepstral coefficients (MFCCs) [3], glottal-based features [4], openSMILE features [5], or sparsity-based features [8]. In addition to SVMs, other classical machine learning methods such as Gaussian Mixture Models [6] and subspace-based learning [9] have been explored. Approaches in the second category have focused on exploring various network architectures and training paradigms such as long short-term memory networks [11], variational autoencoders [12], adversarial training [13], and convolutional neural networks (CNNs) [10, 17]. More recently, self-supervised learning (SSL) methods such as wav2vec2 [18] have been successfully exploited for a variety of speech classification tasks [19], motivating their use for automatic dysarthria classification.

Despite the reported success of automatic dysarthria classification approaches, state-of-the-art literature typically relies on the underlying assumption that recordings from control and dysarthric speakers are obtained in the same noiseless environment using the same recording setup. If recordings for one group of speakers are obtained in a consistently different environment than recordings for the other group of speakers, classifiers trained on such recordings would potentially learn characteristics of the recording environment instead of dysarthric speech characteristics. Unfortunately, such an assumption does not seem to be fulfilled for the commonly used UA-Speech [20] and TORGO [21] databases. Although these databases are made available to the community to develop automatic speech recognition (ASR) systems (where different recording environments and setups can even be desirable in order to develop robust ASR systems), they have also been used to validate a considerable number of state-of-the-art automatic dysarthria classification approaches, such as e.g. [3–5, 11–14, 16].

In this paper, we hypothesize that the reported dysarthria classification results using the UA-Speech and TORGO databases may be reflecting characteristics of the recording environment rather than characteristics of dysarthric speech. To investigate this hypothesis, we first estimate the utterance-level signal-to-noise ratio (SNR) in these databases, confirming the large variability in recording conditions. Further, using voice activity detection (VAD), segments that contain only speech and segments that do not contain any speech are extracted from each utterance in these databases. State-of-the-art dysarthria classification approaches are then trained and validated using only the speech segments or using only the non-speech segments. Remarkably, experimental results show that for both databases, the majority of the considered state-of-the-art approaches achieve the same or even a considerably better dysarthria

This work was supported by the Swiss National Science Foundation project no CRSII5_202228 on "Characterisation of motor speech disorders and processes".

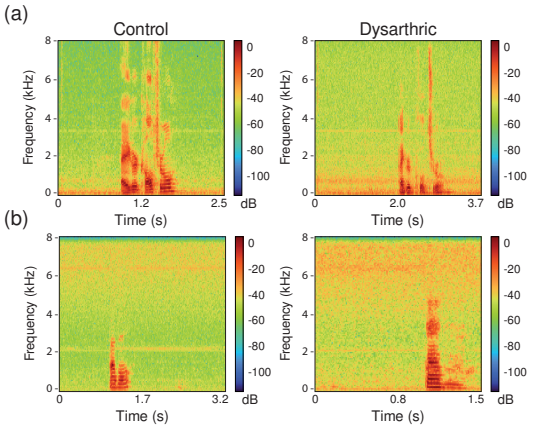

Fig. 1. Spectrograms of an exemplary utterance from a control and dysarthric speaker from the a) UA-Speech and b) TORGO databases.

classification performance when using only non-speech segments than when using only speech segments. These results confirm the hypothesis that dysarthria classification approaches validated on these databases may be learning characteristics of the recording environment rather than dysarthria characteristics.

2. UA-SPEECH AND TORGO DATABASES

In the following, the UA-Speech and TORGO databases are briefly described. Using forced alignment from ASR systems from [22] as VAD, speech segments and non-speech segments are extracted from each utterance in these databases.

UA-Speech [20]. The UA-Speech database contains recordings of 15 patients with CP (4 females, 11 males) and 13 control speakers (4 females, 9 males). Speech signals are sampled at 16 kHz. Since a 7-channel microphone array is used for recording the speakers, we consider the recordings of the 5th-channel (arbitrarily selected) for the evaluations presented in this paper. The number of utterances per speaker is 721 and the average length of all utterances considered for each speaker is 1887 s. Further, the average length of all extracted speech and non-speech segments for each speaker is 564 s and 1323 s respectively.

TORGO [21]. The TORGO database contains recordings from 7 patients (3 females, 4 males) with CP or ALS and from 7 control speakers (3 females, 4 males). Speech signals are sampled at 16 kHz. To avoid additional sources of variability besides dysarthria characteristics, we use only utterances with matched phonetic content across all speakers. The number of such utterances per speaker is 62 and the average length of all utterances considered for each speaker is 201 s. Further, the average length of all extracted speech and non-speech segments for each speaker is 76 s and 125 s respectively.

Fig. 1 depicts spectrograms of exemplary utterances from a control and dysarthric speaker from the UA-Speech and TORGO databases. Visually inspecting the spectrograms in Fig. 1(a) reveals that the exemplary dysarthric spectrogram from the UA-Speech database is noisier than the control spectrogram, particularly at lower frequencies. Further, visually inspecting the spectrograms in Fig. 1(b) reveals that also the exemplary dysarthric spectrogram from the TORGO database is noisier than the control spectrogram, particularly at higher frequencies.

3. METHODS

In this section, the utterance-level SNR estimator is first briefly described. Further, details on the considered state-of-the-art dysarthria classification approaches are presented.

3.1. SNR estimation

Although robust SNR estimation remains an open problem, in this paper we use the recently proposed data-driven recurrent neural network from [23], since it was shown to outperform several state-of-the-art SNR estimators. We use the same network architecture, training procedure, and training and validation datasets as in [23]. The input to the network is the magnitude spectrogram of the noisy signals, whereas the target of the network is the frame-level SNR. Once the frame-level SNR is estimated, an estimate of the utterance-level SNR is obtained as in [23].

3.2. Dysarthria classification approaches

The considered state-of-the-art dysarthria classification approaches are summarized in Table 1. In the following, the handcrafted features or input representations and the classifiers used in these approaches are introduced.

3.2.1. Handcrafted features and input representations

OpenSMILE. As in [5, 7], for each utterance, we extract 6373 features with the openSMILE toolkit [24]. Similarly to [7], dimensionality reduction with Principal Component Analysis is performed by selecting the number of features explaining 95% of the variance in the data (from the training set).

MFCCs. We extract the mean, variance, skewness, and kurtosis of the first 12 MFCCs coefficients using the OpenSMILE toolkit [24], constructing a 48-dimensional feature vector for each utterance.

Sparsity-based features. As in [8], we compute sparsity-based features through the shape parameter of a Chi distribution. To this end, the short-time Fourier transform (STFT) of each utterance is first computed using a Hamming window of length 16 ms and a frame shift of 8 ms. For each frequency bin, a maximum likelihood estimate of the shape parameter of the Chi distribution best modeling the spectral magnitude is obtained. At a sampling frequency of 16 kHz, this procedure yields a 129-dimensional feature vector for each utterance.

Mel spectrograms. Similarly to [15], Mel-scale representations are computed for 500 ms long segments extracted from utterances using a time shift of 250 ms. For each segment, the STFT with a Hamming window of length 32 ms and a frame shift of 4 ms is computed. Final representations are obtained by transforming the STFT coefficients to Mel-scale using 126 Mel bands.

Wav2vec2. Wav2vec2 is a state-of-the-art SSL method that can produce powerful latent speech representations directly from the raw speech signal. The release of the SUPERB benchmark [19] has demonstrated that state-of-the-art results on several speech processing tasks can be achieved by fine-tuning the wav2vec2 model with a lightweight linear prediction classifier. Motivated by these results, in this paper we also exploit the wav2vec2 model for automatic dysarthria classification.

Table 1. Summary of the investigated dysarthria classification approaches.

Approach	Classifier	Handcrafted feature or input representation
SVM+openSMILE [5]	Support vector machine with RBF kernel	ComParE-2016 - openSMILE features
SVM+MFCCs [3]		Mel-frequency cepstral coefficients
SVM+sparsity-based features [8]		Sparsity characterized by the shape parameter
CNN+Mel spectrograms [10]	Convolutional neural network	Mel spectrograms
SRL+Mel spectrograms [15]	Speech representation learning	
MLP+ft-wav2vec2 [19]	Linear classifier	Fine-tuned wav2vec2 embeddings
MLP+wav2vec2 [19]		Wav2vec2 embeddings without fine-tuning

3.2.2. Classifiers

Support vector machines. SVMs are traditional classifiers commonly used with handcrafted acoustic features for dysarthria classification [3–5, 8]. In the following, SVMs with a radial basis kernel function (RBF) are used with different handcrafted acoustic features, i.e., openSMILE, MFCCs, and sparsity-based features.

Convolutional neural networks. CNNs have been widely used to extract discriminative input representations and achieve dysarthric speech classification [10, 14, 17]. In the following, we use a CNN operating on Mel-scale input spectrograms [10]. We adopt the architecture from [10] consisting of two convolutional layers (with 32 and 64 channels, kernel size: 10×10, and stride: 1). Each convolutional layer is followed by batch normalization, max-pooling (kernel size: 2, stride: 3), and the ReLU activation function. A dropout layer with a rate of 20% is placed after the final convolutional layer. A final fully-connected layer with 128 input units and 2 output units is used for dysarthria classification.

Speech representation learning (SRL). In this paper, SRL is used to refer to the state-of-the-art dysarthria classification approach proposed in [15], where a CNN-based auto-encoder is used to learn low dimensional discriminative bottleneck representations from Mel-scale input spectrograms. Bottleneck representations are learned by jointly minimizing the auto-encoder loss and the loss of a linear dysarthria classifier. The learned representations are then fine-tuned for the final dysarthria classification network. The architecture description of the network can be found in [15].

Multilayer perceptron (MLP). Motivated by [19], we also evaluate the performance of an MLP trained on wav2vec2 embeddings for dysarthria classification. The MLP consists of two fully-connected layers. The first layer has 768 input units and 256 output units and the second layer has 256 input units and 2 output units. Similarly to the speaker identification task in [19], each utterance is processed by the wav2vec2 model and the obtained embeddings are mean-pooled prior to being forwarded to the MLP classifier. As outlined in Table 1 and as described in Section 4.1, we consider two approaches using wav2vec2, i.e., *MLP+ft-wav2vec2* referring to fine-tuning parts of the wav2vec2 model together with the MLP for dysarthria classification and *MLP+wav2vec2* referring to freezing the wav2vec2 model and training only the MLP.

4. EXPERIMENTAL RESULTS

In this section, the utterance-level SNR of control and dysarthric recordings from the UA-Speech and TORGO databases is analyzed. Further, the performance of state-of-the-art dysarthria classification approaches when using only speech segments and only non-speech segments from these databases is analyzed. For completeness, the performance of the considered classification approaches when using the complete utterances without any VAD (i.e., both speech and non-speech segments) is also presented.

4.1. Training and validation

For all approaches investigated in this paper (cf. Table 1), a leave-one-speaker-out validation strategy is used. In each fold, 90% of the data from the training speakers is used for training, whereas 10% of the data is used for validation. The prediction for a test speaker is made through majority voting of the utterance-level/segment-level predictions and the final performance is evaluated in terms of the speaker-level classification accuracy. To reduce the impact that random initialization has on the final performance, we have trained all approaches using 3 different random initialization. The reported final performance for all approaches is the mean and standard deviation of the speaker-level classification accuracy obtained across these different models. Except for the wav2vec2 embeddings, we apply z-score standardization to all handcrafted acoustic features and input representations. In the following, details on the training of each considered approach are presented.

SVMs. Separate SVMs are trained for each handcrafted acoustic feature in Table 1. The soft margin constant C and the kernel width γ are optimized using a grid search procedure with $C \in \{10, 10^4\}$ and $\gamma \in \{10^{-4}, 10^{-1}\}$. The optimal hyperparameters are selected as the ones that achieve the highest utterance-level classification accuracy on the validation set.

CNN+Mel spectrograms. The CNN is trained using the Adam optimizer and the cross–entropy loss function. We use a batch size of 128 and an initial learning rate of 2×10^{-5} for a total of 50 epochs. A scheduler is set to halve the learning rate if the loss on the validation set does not decrease for 5 consecutive iterations.

SRL+Mel spectrograms. As in [15], the stochastic gradient descent algorithm is used for training the SRL approach. The dysarthria classifier is trained using cross-entropy, whereas the auto-encoder is trained using mean square error. Further, we use a batch size of 128 and an initial learning rate of 0.02 for a total of 20 epochs. A scheduler is set to halve the learning rate if the loss on the validation set does not decrease for 5 consecutive iterations.

MLP+ft-wav2vec2. To fine-tune the wav2vec2 model, we freeze the CNN encoder and fine-tune the transformer and the MLP classifier. As in [19], the AdamW optimizer and the cross-entropy loss function are used. Training is done with an effective batch size of 128, i.e., a batch size of 16 and a gradient accumulation step of 8. A linear warm-up scheduler is used (warm-up ratio: 0.1) and the initial learning rate is set to 3×10^{-5}.

MLP+wav2vec2. Using the wav2vec2 embeddings without fine-tuning refers to freezing the complete wav2vec2 model and training only the MLP classifier for dysarthria classification. The

Table 2. Mean and standard deviation of the estimated SNRs [dB] across all utterances of control and dysarthric speakers in the UA-Speech and TORGO databases.

Speakers	UA-Speech	TORGO
Control	3.7 ± 11.5	2.1 ± 13.2
Dysarthric	-7.6 ± 16.1	-4.0 ± 14.7

Table 3. Mean and standard deviation of the speaker classification accuracy [%] across all folds and models in the UA-Speech database.

Approach	Speech	Non-speech	Speech&Non-speech
SVM+openSMILE	81.0 ± 19.8	84.5 ± 21.9	83.3 ± 21.1
SVM+MFCCs	81.0 ± 1.7	100.0 ± 0.0	100.0 ± 0.0
SVM+sparsity-based features	94.0 ± 1.7	96.4 ± 0.0	96.4 ± 0.0
CNN+Mel spectrograms	95.2 ± 1.7	97.6 ± 1.7	98.8 ± 1.7
SRL+Mel spectrograms	98.8 ± 1.7	100.0 ± 0.0	100.0 ± 0.0
MLP+ft-wav2vec2	95.2 ± 1.7	97.6 ± 1.7	95.2 ± 1.7
MLP+wav2vec2	54.8 ± 1.7	58.3 ± 1.7	54.8 ± 1.7

used optimizer, loss function, batch size, and learning rate are the same as for the *MLP+ft-wav2vec2* approach.

4.2. Results

SNR estimation. Table 2 presents the mean and standard deviation of the estimated utterance-level SNRs across all control and dysarthric utterances for the UA-Speech and TORGO databases. As demonstrated by the large standard deviation values of the estimated SNRs, it can be said that there is a large variation in the acoustic conditions of the recorded utterances for both databases. Most importantly, it can be observed that there is a large difference in the average SNRs of control and dysarthric utterances in both databases, with the difference being larger for the UA-Speech database[1]. With consistently different recording conditions between control and dysarthric utterances, there is no guarantee that automatic dysarthria classification approaches validated on these databases are learning control and dysarthric speech differences instead of differences in recording conditions for the two groups of speakers.

Dysarthria classification. Table 3 presents the mean and standard deviation of the classification accuracy obtained on the speech segments, the non-speech segments, and on the complete utterances without using any VAD (i.e., speech&non-speech) for the UA-Speech database using all considered approaches (cf. Table 1). It can be observed that all approaches achieve the same or even better dysarthria classification accuracy when using non-speech segments in comparison to when using speech segments or the complete speech&non-speech segments. More specifically, it can be observed that when using non-speech segments, all approaches except for *MLP+wav2vec2* yield a high classification accuracy ranging from 84.5% to 100.0%. The *MLP+wav2vec2* approach performs considerably worse than its fine-tuned version *MLP+ft-wav2vec2* and all other considered approaches. This result is to be expected since the representations generated by the (frozen) wav2vec2 model should be less susceptible to noise given that the model is trained on a large database of noisy speech.

[1] Although not presented here due to space constraints, the utterance-level SNRs have been estimated using different SNR estimators. While the absolute value of the estimated SNRs can be largely different depending on the used SNR estimator, all estimators show large standard deviation values and considerable differences between the average SNRs of control and dysarthric utterances in both databases.

Table 4. Mean and standard deviation of the speaker classification accuracy [%] across all folds and models in the TORGO database.

Approach	Speech	Non-speech	Speech&Non-speech
SVM+openSMILE	60.0 ± 5.4	82.2 ± 6.3	71.1 ± 12.6
SVM+MFCCs	60.0 ± 0.0	88.9 ± 3.1	57.8 ± 3.1
SVM+sparsity-based features	73.3 ± 0.0	93.3 ± 0.0	73.3 ± 5.4
CNN+Mel spectrograms	53.3 ± 11.5	77.8 ± 10.2	68.9 ± 10.2
SRL+Mel spectrograms	71.1 ± 3.1	100.0 ± 0.0	91.1 ± 3.1
MLP+ft-wav2vec2	60.0 ± 5.4	57.8 ± 3.1	60.0 ± 5.4
MLP+wav2vec2	55.6 ± 3.1	57.8 ± 3.1	57.8 ± 6.3

Table 4 presents the mean and standard deviation of the classification accuracy obtained on the speech, non-speech, and the complete speech and non-speech segments from the TORGO database using all considered approaches (cf. Table 1). Similarly to before, it can be observed that all approaches achieve the same or even better dysarthria classification accuracy when using non-speech segments in comparison to when using speech segments or the complete speech&non-speech segments. Further, it can be observed that the *MLP+wav2vec2* approach is not as sensitive to the recording conditions as the other approaches, as illustrated by the lower performance on non-speech segments. However, differently from before, the performance of the fine-tuned counterpart *MLP+ft-wav2vec2* on non-speech segments is also low. We suspect this occurs due to the much smaller amount of speech material available for fine-tuning the wav2vec2 model on the TORGO database (in contrast to the UA-Speech database).

In summary, the results presented in this section show that the majority of the considered state-of-the-art approaches achieve the same or even better dysarthria classification performance when using non-speech segments than when using speech segments or complete utterances from the UA-Speech and TORGO databases. These results confirm our hypothesis that classification results obtained on the UA-Speech and TORGO databases can be greatly affected by characteristics of the recording environment and setup instead of dysarthria characteristics.

5. CONCLUSIONS

In this paper, we have investigated the use of the UA-Speech and TORGO databases to validate automatic dysarthria classification approaches. We hypothesized that classification results obtained using these databases could be biased towards capturing characteristics of the recording environment rather than characteristics of dysarthric speech. To investigate this hypothesis, we have estimated the utterance-level SNRs on these databases. Further, we have trained and validated state-of-the-art dysarthria classification approaches on the speech and non-speech segments of these databases. Experimental results have shown that the utterance-level SNRs in control and dysarthric recordings are indeed considerably different in both databases. Additionally confirming our hypothesis, experimental results have shown that several state-of-the-art approaches achieve the same or a considerably better dysarthria classification performance when using only the non-speech segments than when using only the speech segments. We hope that these results raise awareness in the research community about the care that should be taken with respect to the quality of recordings when developing and evaluating automatic dysarthria classification approaches.

6. REFERENCES

[1] F. L. Darley, A. E. Aronson, and J. R. Brown, "Differential diagnostic patterns of dysarthria," *Journal of Speech and Hearing Research*, vol. 12, pp. 246–269, June 1969.

[2] R. D. Kent, "Hearing and believing: Some limits to the auditory-perceptual assessment of speech and voice disorders," *American Journal of Speech-Language Pathology*, vol. 5, no. 3, pp. 7–23, Aug. 1996.

[3] K. L. Kadi, S. A. Selouani, B. Boudraa, and M. Boudraa, "Fully automated speaker identification and intelligibility assessment in dysarthria disease using auditory knowledge," *Biocybernetics and Biomedical Engineering*, vol. 36, no. 1, pp. 233–247, Nov. 2016.

[4] S. Gillespie, Y. Logan, E. Moore, J. Laures-Gore, S. Russell, and R. Patel, "Cross-database models for the classification of dysarthria presence," in *Proc. Annual Conference of the International Speech Communication Association*, Stockholm, Sweden, Aug. 2017, pp. 3127–3131.

[5] N. P. 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 Association*, Hyderabad, India, Sept. 2018, pp. 3403–3407.

[6] L. Jeancolas, G. Mangone, J. Corvol, M. Vidailhet, S. Lehericy, B. Benkelfat, H. Benali, and D. Petrovska-Delacretaz, "Comparison of telephone recordings and professional microphone recordings for early detection of Parkinson's disease, using mel-frequency cepstral coefficients with Gaussian mixture models," in *Proc. Annual Conference of the International Speech Communication Association*, Graz, Austria, Sept. 2019, pp. 3033–3037.

[7] 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.

[8] 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, Apr. 2020.

[9] 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.

[10] 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 Association*, Stockholm, Sweden, Aug. 2017, pp. 314–318.

[11] J. Millet and 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.

[12] J. Qi and H. Van Hamme, "Speech disorder classification using extended factorized hierarchical variational auto-encoders," in *Proc. Annual Conference of the International Speech Communication Association*, Brno, Czechia, Aug. 2021, pp. 1917–1921.

[13] D. Wang, L. Deng, Y. T. Yeung, X. Chen, X. Liu, and H. Meng, "Unsupervised domain adaptation for dysarthric speech detection via domain adversarial training and mutual information minimization," in *Proc. Annual Conference of the International Speech Communication Association*, Brno, Czechia, Aug. 2021, pp. 2956–2960.

[14] 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.

[15] 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.

[16] 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.

[17] 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, Singapore, May 2022, pp. 6477–6481.

[18] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. International Conference on Neural Information Processing Systems*, Vancouver, Canada, Dec. 2020, pp. 12449–12460.

[19] S. Yang, P. Chi, Y. Chuang, C. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G. Lin, et al., "Superb: Speech processing universal performance benchmark," in *Proc. Annual Conference of the International Speech Communication Association*, Brno, Czechia, Aug. 2021, pp. 1194–1198.

[20] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. S. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," in *Proc. Annual Conference of the International Speech Communication Association*, Brisbane, Australia, Sept. 2008, pp. 1741–1744.

[21] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, vol. 46, pp. 523–541, Mar. 2012.

[22] E. Hermann and M. Magimai-Doss, "Handling acoustic variation in dysarthric speech recognition systems through model combination.," in *Proc. Annual Conference of the International Speech Communication Association*, Brno, Czechia, Aug. 2021, pp. 4788–4792.

[23] H. Li, D. Wang, X. Zhang, and G. Gao, "Recurrent neural networks and acoustic features for frame-level signal-to-noise ratio estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2878–2887, Aug. 2021.

[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*, Firenze, Italy, Oct. 2010, pp. 1459–1462.