Adversarial-Free Speaker Identity-Invariant Representation Learning for Automatic Dysarthric Speech Classification

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

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

Abstract

Speech representations which are robust to pathology-unrelated cues such as speaker identity information have been shown to be advantageous for automatic dysarthric speech classification. A recently proposed technique to learn speaker identity-invariant representations for dysarthric speech classification is based on adversarial training. However, adversarial training can be challenging, unstable, and sensitive to training parameters. To avoid adversarial training, in this paper we propose to learn speaker-identity invariant representations exploiting a feature separation framework relying on mutual information minimization. Experimental results on a database of neurotypical and dysarthric speech show that the proposed adversarial-free framework successfully learns speaker identity-invariant representations. Further, it is shown that such representations result in a similar dysarthric speech classification performance as the representations obtained using adversarial training, while the training procedure is more stable and less sensitive to training parameters.
Index Terms: Parkinson's disease, speaker identity, feature separation, supervised autoencoder, mutual information

1. Introduction

Neurodegenerative disorders such as Parkinson's disease (PD) can cause speech dysarthria, resulting in impairments in the speech production mechanism and reduced communicative ability [1]. To assist the clinical diagnosis of dysarthria, automatic machine learning techniques have been developed. Such techniques provide reliable, objective, and cost-effective assessments in contrast to the subjective and time-consuming auditory-perceptual analyses performed by clinicians [2].

The majority of state-of-the-art automatic dysarthric speech classification techniques are based on training classical classifiers on handcrafted acoustic features characterizing different impaired speech dimensions [3–9]. Recently, deep learning approaches aiming to learn high-level speech representations relevant for such a task have gained attention in the research community [10–18]. Due to the large number of parameters in the used networks and the small amount of pathological training data that is typically available, deep learning approaches for automatic dysarthric speech classification can be sensitive to pathology-unrelated variabilities such as speaker identity cues [19]. To increase the number of training samples, state-of-the-art techniques commonly split available utterances into many short segments and individually classify each segment as healthy or dysarthric using convolutional neural networks (CNNs) [10–15]. However, such short segments do not necessarily contain dysarthric cues and the used CNNs are not guided to ignore speaker variabilities that are unrelated to dysarthria. To cope with the small number of available training utterances while extracting more robust representations, we proposed using a CNN operating on pairwise distance matrices constructed from complete utterances [20]. Although advantageous, such an approach relies on having access to utterances with the same phonetic content from both healthy and pathological speakers. To relax these phonetic constraints while explicitly learning a representation that is robust to pathology-unrelated cues such as speaker identity information, we recently proposed to obtain speaker identity-invariant representations on short (phonetically-unmatched) speech segments using a supervised representation learning framework [19]. To this end, we exploited a supervised autoencoder (AE) with an adversarial speaker identification (ID) module to learn bottleneck representations containing no speaker identity cues. Such representations were then used as input to a neural network for dysarthria classification. We showed that using speaker-identity invariant representations for dysarthria classification yields a significantly better performance than unsupervised representations containing speaker identity cues [19]. It should be noted that besides improving the dysarthria classification performance, another important motivation for suppressing speaker identity cues arises in the context of voice privacy preservation [21]. As outlined in the recently organized VoicePrivacy challenge in [22], recent years have seen increasing pressure for privacy-preserving speech technologies suppressing speaker information from speech representations, while preserving the paralinguistic acoustic cues related to pathological conditions.

The adversarial training framework used in [19] to obtain speaker-identity invariant representations is challenging, since it can be very sensitive to training parameters and it can result in oscillating, unstable, and divergent models [23]. Furthermore, it has been shown that adversarial training can be unnecessary and counter-productive [24]. To avoid adversarial training, in this paper we propose an adversarial-free framework to obtain speaker identity-invariant representations using feature separation through mutual information (MI) minimization. The proposed adversarial-free framework consists of two encoders and one decoder. The first encoder generates a bottleneck representation containing speaker identity cues, whereas the second encoder generates a speaker identity-invariant bottleneck representation. To ensure the presence of speaker identity cues in the first bottleneck representation (generated by the first encoder), the performance of a speaker ID auxiliary classifier operating on this representation is maximized. To reduce the presence of speaker identity cues in the second bottleneck representation (generated by the second encoder), the MI between the two bottleneck representations is minimized. To avoid information loss, a decoder fed by both encoded representations is simultaneously trained to minimize the reconstruction loss. Such a training procedure avoids adversarial training and yields a representation (generated by the second encoder) with suppressed

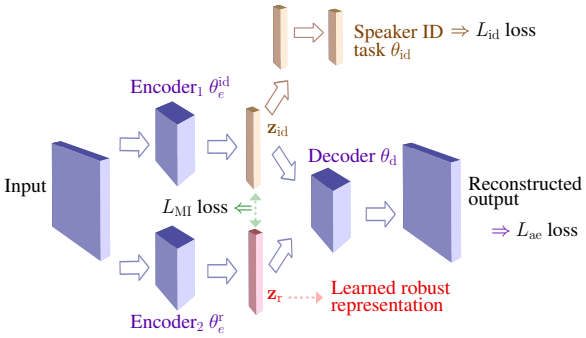


Figure 1: *Proposed feature separation framework.*

speaker identity cues, which is consequently a more robust representation for dysarthric speech classification.

MI minimization to separate latent representations into (ideally) independent components while avoiding adversarial training has been exploited for several applications, e.g., unsupervised domain adaptation for image classification tasks, voice style transfer (voice conversion), and speech synthesis [25–27]. As estimating the MI between high dimensional continuous variables is a difficult task, different estimators of the upper and lower bounds of MI using neural network architectures have been proposed and considered for optimization instead [25, 28]. To minimize the MI in the proposed adversarial-free framework, we use the neural network-based MI estimator in [25].

Experimental results on a Spanish database of neurotypical and PD speakers show that the proposed adversarial-free framework successfully learns a speaker identity-invariant representation, while being more robust than the adversarial framework to the choice of training parameters. Further, it is shown that using this representation for dysarthric speech classification yields a substantially better performance than unsupervised representations containing speaker identity cues.

2. Adversarial-Free Representation Learning for Dysarthria Classification

To suppress speaker identity cues while avoiding adversarial training, we propose to use the feature separation framework schematically illustrated in Figure 1. In this framework, chunks of time-frequency input representations are projected onto a pair of bottleneck representations using two encoders. A decoder is simultaneously trained to generate a reconstructed version of the input using the concatenated bottleneck representations to avoid information loss. The first bottleneck representation $\mathbf{z}_{\text{id}}$ should only contain information about speaker identities, while the second (residual) bottleneck representation $\mathbf{z}_{\text{r}}$ should contain no information about speaker identities. To this end, $\mathbf{z}_{\text{id}}$ is directly supervised by a speaker ID classifier, while $\mathbf{z}_{\text{r}}$ is separated from $\mathbf{z}_{\text{id}}$ by minimizing a MI criterion between them. Since all modules, i.e., the two encoders, the decoder, and the auxiliary speaker ID classifier, are jointly trained, it is expected that the bottleneck representation $\mathbf{z}_{\text{r}}$ encodes speaker identity-invariant cues, making $\mathbf{z}_{\text{r}}$ a more robust representation for dysarthria classification. In the following, the architecture and training procedure for the proposed speaker identity-invariant representation learning are presented.

2.1. Modules

Adapted from [19], each encoder contains three convolutional layers (filter size: 6×6, stride: 1), with the number of feature maps on each layer being twice the number of feature maps on the previous layer (starting with 32 maps in the first layer). Each convolutional layer is followed by max-pooling (filter size: 3×3, stride: 3), batch normalization, and a ReLU activation function. The output of the last convolutional layer is further processed by a fully connected layer (with 128 hidden units) to form the bottleneck representation of dimension 128. The architecture of both encoders in the framework is the same. The outputs of the two encoders are concatenated and fed to a decoder. The decoder components are stacked in reverse order of the encoder components, where transposed convolutional and interpolation layers are used instead of convolutional and max-pooling layers. The parameters of the two encoders generating the representations $\mathbf{z}_{\text{id}}$ and $\mathbf{z}_{\text{r}}$ are denoted by θ_e^{id} and θ_e^{r}, respectively, and the decoder parameters are denoted by θ_{d}. For the speaker ID classifier, the same architecture as in [19] is used. The number of output units, i.e., the number of units in the final layer of the speaker ID module, is equal to the number of speakers used for the speaker ID task (cf. Section 4.2). The parameters of this module are denoted by θ_{id}. As described in the following, the proposed framework also consists of an MI estimator module which is needed for estimating and minimizing the MI between $\mathbf{z}_{\text{id}}$ and $\mathbf{z}_{\text{r}}$.

2.2. MI minimizer

The MI $I(\mathbf{z}_{\text{id}}, \mathbf{z}_{\text{r}})$ between $\mathbf{z}_{\text{id}}$ and $\mathbf{z}_{\text{r}}$ is defined as the Kullback-Leibler divergence between the joint distribution and the product of marginal distributions of the two variables, i.e., $I(\mathbf{z}_{\text{id}}, \mathbf{z}_{\text{r}}) = D_{KL}\big(p(\mathbf{z}_{\text{id}}, \mathbf{z}_{\text{r}}) || p(\mathbf{z}_{\text{id}})p(\mathbf{z}_{\text{r}})\big)$. Since MI computation is challenging for high-dimensional variables with unknown probability distributions, variational contrastive log-ratio upper bound (vCLUB) is used in [25] as an upper bound for MI, i.e.,

$$\begin{aligned} I^{\phi}_{\text{vCLUB}}(\mathbf{z}_{\text{id}}, \mathbf{z}_{\text{r}}) =& \mathbb{E}_{p(\mathbf{z}_{\text{id}}, \mathbf{z}_{\text{r}})}\left[\log q_\phi(\mathbf{z}_{\text{id}}|\mathbf{z}_{\text{r}})\right] \\ &- \mathbb{E}_{p(\mathbf{z}_{\text{id}})}\mathbb{E}_{p(\mathbf{z}_{\text{r}})}\left[\log q_\phi(\mathbf{z}_{\text{id}}|\mathbf{z}_{\text{r}})\right], \end{aligned} \tag{1}$$

where $q_\phi(\mathbf{z}_{\text{id}}|\mathbf{z}_{\text{r}})$ is the Gaussian variational approximation of $p(\mathbf{z}_{\text{id}}|\mathbf{z}_{\text{r}})$ with mean $\mu(\mathbf{z}_{\text{r}})$ and variance $\sigma^2(\mathbf{z}_{\text{r}})$. The mean and variance estimation is done through neural networks with parameters denoted by ϕ. The networks consist of a fully connected layer with 64 hidden units, a ReLU activation function, and a 128–dimensional output vector representing $\mu(\mathbf{z}_{\text{r}})$ or $\sigma^2(\mathbf{z}_{\text{r}})$ [25]. For the variance estimating network, a Tanh (hyperbolic tangent) activation function is applied after the output. The network parameters ϕ are approximated by maximizing the log-likelihood loss $L_{\text{ll}}(\phi) = \log q_\phi(\mathbf{z}_{\text{id}}|\mathbf{z}_{\text{r}})$ as in [25]. After obtaining an estimate $\hat{\phi}$ of the network parameters, we use the vCLUB as our MI objective to be minimized, i.e., $L_{\text{MI}}(\theta_e^{\text{id}}, \theta_e^{\text{r}}, \hat{\phi}) = I^{\hat{\phi}}_{\text{vCLUB}}(\mathbf{z}_{\text{id}}, \mathbf{z}_{\text{r}})$.

2.3. Feature separation

Learning the speaker identity-invariant representation $\mathbf{z}_{\text{r}}$ is achieved through minimizing the objective function

$$\begin{aligned} E(\theta_e^{\text{id}}, \theta_e^{\text{r}}, \theta_{\text{d}}, \theta_{\text{id}}, \phi) =& L_{\text{ae}}(\theta_e^{\text{id}}, \theta_e^{\text{r}}, \theta_{\text{d}}) + \lambda L_{\text{id}}(\theta_e^{\text{id}}, \theta_{\text{id}}) \\ &+ \beta L_{\text{MI}}(\theta_e^{\text{id}}, \theta_e^{\text{r}}, \phi), \end{aligned} \tag{2}$$

with L_{ae} and L_{id} being the AE reconstruction and speaker ID loss functions, respectively, and λ and β being the weights of the speaker ID and MI loss functions (cf. Section 4.2). Be-

cause of the MI estimator module, optimal parameters in (2) are approximated using an alternating training procedure, i.e.,

$$(\hat{\theta}_e^{\mathrm{id}}, \hat{\theta}_e^{\mathrm{r}}, \hat{\theta}_{\mathrm{d}}, \hat{\theta}_{\mathrm{id}}) = \arg \min_{\theta_e^{\mathrm{id}}, \theta_e^{\mathrm{r}}, \theta_{\mathrm{d}}, \theta_{\mathrm{id}}} E(\theta_e^{\mathrm{id}}, \theta_e^{\mathrm{r}}, \theta_{\mathrm{d}}, \theta_{\mathrm{id}}, \hat{\phi}), \quad (3)$$

$$\hat{\phi} = \arg \min_{\phi} -L_{\mathrm{ll}}(\phi, \hat{\theta}_e^{\mathrm{id}}, \hat{\theta}_e^{\mathrm{r}}). \quad (4)$$

2.4. Dysarthric Speech Classification

As in [19], the learned speaker identity-invariant representation $\mathbf{z}_{\mathrm{r}}$ is used as input for a dysarthric speech classifier. The architecture of this classifier is identical to the speaker ID classifier in Section 2.1, except for the number of output units being 2 (since we are dealing with binary classification, i.e., dysarthric vs. neurotypical speech). The final decision for an unseen (test) speaker is made by averaging the classifier prediction scores for all input representations belonging to that speaker.

3. Adversarial Representation Learning for Dysarthria Classification

For completeness, in the following we briefly describe the adversarial training framework for learning speaker identity-invariant representations from [19].

Considering the schematic representation in Figure 1, the adversarial training framework consists only of the first encoder $\theta_{\mathrm{e}}^{\mathrm{id}}$, the decoder θ_{d}, and the speaker ID module. The architecture of these modules is as described in Section 2.1, with the only difference being the size of the first decoder layer (since differently from the proposed feature separation framework, only one bottleneck representation is encoded and decoded in the adversarial framework).[1] To obtain an encoded representation where speaker identity cues are suppressed, a gradient reversal layer is included before the speaker ID module. Hence, the adversarial training optimization objective consists of only the AE reconstruction and speaker ID loss, with a sign reversal for the speaker ID loss, i.e.,

$$E_{\mathrm{adv}}(\theta_e^{\mathrm{id}}, \theta_{\mathrm{d}}, \theta_{\mathrm{id}}) = L_{\mathrm{ae}}(\theta_e^{\mathrm{id}}, \theta_{\mathrm{d}}) - \lambda L_{\mathrm{id}}(\theta_e^{\mathrm{id}}, \theta_{\mathrm{id}}). \quad (5)$$

Optimization of (5) is done in an alternating fashion, where in the first step, the AE parameters θ_e^{id} and θ_{d} are updated assuming fixed speaker ID parameters θ_{id}, and in the second step, the parameters θ_{id} are updated assuming fixed parameters θ_e^{id} and θ_{d} obtained in the first step.

4. Experimental Results

In this section, the performance of dysarthric speech classification using the proposed adversarial-free representation learning framework is evaluated and compared to using the adversarial representation learning framework from [19]. Furthermore, the efficacy of the speaker ID and MI minimizer modules in the proposed framework is also investigated. Empirical insights regarding the stability of model training with respect to several training parameters are also provided.

[1] It should be noted that the architecture of the adversarial training framework used here differs from the one in [19] such that the adversarial and adversarial-free training frameworks can be fairly compared under the same architecture.

4.1. Database

We consider Spanish recordings from 50 PD patients (25 males, 25 females) and 50 neurotypical speakers (25 males, 25 females) from the PC-GITA database [29]. Each speaker utters 24 words, 10 sentences, and 1 text recorded at a sampling frequency of 44.1 kHz. After downsampling to 16 kHz, speech-only segments are manually extracted from the word recordings and using an energy-based voice activity detector for all other recordings [30]. The average length of the speech material considered for each speaker is 59.9 s.

4.2. Training, evaluation, and baseline systems

Training. Representations are learned from Mel-scale input representations of 500 ms segments of speech with the same settings as in [19]. For training and evaluation, we use a stratified speaker-independent 10-fold cross-validation. In each training fold, a development fold of the same size as the test fold is set aside for early-stopping when training the final dysarthric speech classifier. For the speaker ID module, utterances from the healthy speakers in the training set (i.e., 45 speakers) are split without overlap into 50% train, 25% development, and 25% test sets. Cross-entropy is used for the speaker ID loss L_{id} and for the final dysarthric speech classifier, whereas the mean square error of the reconstructed representation is used for the AE loss L_{ae}.

All parameters are estimated using the ADAM optimizer [31]. Preliminary results showed that the learning rate for each module of the considered frameworks should be different. Using different learning rates is particularly important for the adversarial training framework. For the results presented in the following, learning rates are empirically set to 10^{-5} for the AE module, 10^{-3} for the speaker ID classifier module, and 10^{-3} for the MI estimator network. The representation learning frameworks are trained with a batch size of 128 for 50 epochs. The final dysarthric speech classifier is trained using a learning rate of 10^{-4} after freezing the encoder parameters. This learning rate is halved each time the classification loss on the development set does not decrease for 5 consecutive iterations. Training is stopped either after 50 epochs or after the classifier learning rate has decreased beyond 0.1 of the initial learning rate. To investigate the suppression of speaker identity cues in the learned representations, a speaker ID classifier (with the same architecture as the speaker ID module in Section 2.1) is also trained with an initial learning rate of 10^{-3} following the same early-stopping procedure described above.

Evaluation. Dysarthric speech classification performance is evaluated in terms of accuracy (i.e., percentage of correctly classified neurotypical and PD speakers) and the AUC. The performance of the speaker ID classifier is also evaluated in terms of accuracy (i.e., percentage of correctly identified speakers) and AUC. To reduce the impact of random initialization, all networks are trained with 5 different random seeds. The performance values reported in the following are the mean and standard deviation of the performance across different seeds.

Baseline systems. We compare the representation obtained by the proposed adversarial-free framework to the representation obtained via adversarial training from [19].[2] Comparisons are done in terms of dysarthric speech classification performance as well as speaker ID performance. An advantageous

[2] Note that experimental results demonstrating the advantages of suppressing speaker identity cues as opposed to state-of-the-art approaches are provided in [19].

representation should result in a low speaker ID performance (i.e., meaning that speaker ID cues are suppressed) and a high dysarthric speech classification performance. The adversarial-free and adversarial representation learning frameworks require setting the loss weights λ and β (cf. (2) and (5)). In the following, these weights are set using grid-search over a set of 8 values between 10^{-4} and 0.5, with the final weights for each framework selected as the ones yielding the highest mean dysarthric speech classification accuracy on the development set. To investigate the effects of the auxiliary modules in the proposed framework, we also consider baseline systems generated by excluding the supervision of these modules. Without using the speaker ID and MI minimizer modules in training, i.e., setting $\lambda = \beta = 0$ in (2), we obtain an unsupervised baseline system. Keeping the speaker ID module while removing the MI minimizer during training, i.e., setting $\beta = 0$ in (2), we obtain a partially supervised baseline system.

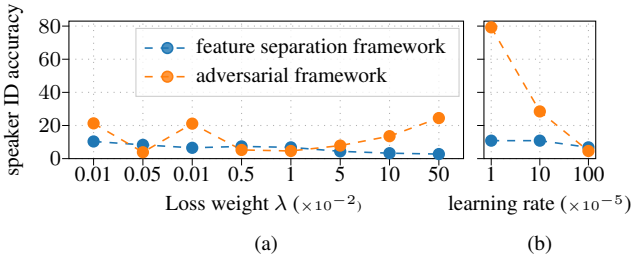
4.3. Results

Table 1 presents the performance values obtained using representations learned from the unsupervised baseline framework (i.e., $\lambda = \beta = 0$), partially supervised baseline framework (i.e., $\beta = 0$), proposed adversarial-free framework, and state-of-the-art adversarial framework. First, it can be observed that using the representations from the unsupervised and partially supervised baselines, a relatively high speaker ID performance and a low dysarthric speech classification performance is achieved. This is to be expected since in both models, no supervision is used for isolating speaker identity cues from $\mathbf{z}_{\text{r}}$, making $\mathbf{z}_{\text{r}}$ an unrobust representation for dysarthria classification. Second, it can be observed that the representation learned by the proposed method (where both speaker ID and MI minimizer modules are included in training) gives a substantially higher dysarthric speech classification performance and a substantially lower speaker ID performance. These results confirm the efficacy of the proposed adversarial-free framework to obtain a speaker identity-invariant representation, and therefore, a more robust representation for dysarthria classification. Third, comparing the results of the proposed adversarial-free framework and the state-of-the-art adversarial framework, it can be observed that the dysarthria classification and speaker ID performance values are similar. Statistical significance analysis conducted as in [7] using a corrected resampled t-test show that the slight difference in performance between the two frameworks is not statistically significant for a considered threshold of 0.05.

To demonstrate the training advantages of the proposed

Table 1: *Mean and standard deviation of dysarthric speech classification and speaker ID accuracy and AUC values obtained using differently learned representations.*

λ	β	PD classification ↑		speaker ID ↓	
		Accuracy (%)	AUC	Accuracy (%)	AUC
✗	✗	57.2 ± 5.4	0.72 ± 0.02	58.3 ± 2.1	0.98 ± 0.00
✓	✗	61.4 ± 3.4	0.75 ± 0.02	49.6 ± 3.2	0.98 ± 0.00
Proposed adversarial-free feature separation framework					
✓	✓	75.2 ± 3.5	0.82 ± 0.03	5.0 ± 5.2	0.67 ± 0.09
Adversarial framework from [19]					
–	–	77.0 ± 4.2	0.85 ± 0.03	5.2 ± 2.2	0.67 ± 0.05



(a) (b)

Figure 2: *Speaker ID accuracy of the speaker identity-invariant representations obtained from the proposed feature separation and the adversarial training in [19] for different a) loss weight values and b) learning rates of speaker ID module.*

adversarial-free framework as opposed to the adversarial framework, in the following we investigate the suppression of speaker ID cues in the learned representations as a function of two different training parameters, i.e., loss weight λ and learning rate of the speaker ID module (cf. (2) and (5)). For the proposed adversarial-free framework, we use $\beta = 0.5$. The suppression of speaker ID cues is evaluated through the speaker ID accuracy, with low accuracy values implying a high suppression and vice-versa. Figure 2a shows the speaker ID accuracy obtained from the representations of both frameworks for different loss weights λ. It can be observed that compared to adversarial training, the suppression of speaker ID cues in the proposed framework is less sensitive to λ. Figure 2b shows the speaker ID accuracy obtained from the representations of both frameworks for different learning rates of the speaker ID module and $\lambda = 0.01$. It can be observed that compared to adversarial training, the adversarial-free framework achieves a low speaker ID accuracy independently of the learning rate of the speaker ID module. These results confirm that the suppression of speaker identity cues in the proposed adversarial-free framework is less sensitive to training parameters than in the adversarial framework.

In summary, the presented results show that the proposed adversarial-free training framework successfully learns a speaker identity-invariant representation which is advantageous for dysarthric speech classification, while being less sensitive to training parameters than the existing adversarial training framework.

5. Conclusion

In this paper, we have proposed a supervised representation learning framework for dysarthric speech classification. To suppress speaker identity cues unrelated to dysarthria, we have exploited an adversarial-free feature separation framework based on training a dual encoder and a single decoder. To enforce speaker identity cues to be present only in one of the encoded representations, we have supervised one of the representations with a speaker ID auxiliary task while minimizing a MI criterion between the two representations. Experimental results have shown that the proposed framework is successful in learning a speaker identity-invariant representation, while being more robust to training parameters when compared to the state-of-the-art adversarial framework.

6. Acknowledgments

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

7. References

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

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

[3] J. R. Orozco-Arroyave, F. Honig, J. D. Arias-Londono, J. F. Vargas-Bonilla, K. Daqrouq, S. Skodda, J. Rusz, and E. Noeth, "Automatic detection of Parkinson's disease in running speech spoken in three different languages," *The Journal of the Acoustical Society of America*, vol. 139, no. 1, pp. 481–500, Jan. 2016.

[4] I. Kodrasi and H. Bourlard, "Super-Gaussianity of speech spectral coefficients as a potential biomarker for dysarthric speech detection," in *Proc. 44th IEEE International Conference on Acoustics, Speech and Signal Processing*, Brighton, UK, May 2019, pp. 6400–6404.

[5] G. Solana-Lavalle and R. Rosas-Romero, "Analysis of voice as an assisting tool for detection of parkinson's disease and its subsequent clinical interpretation," *Biomedical Signal Processing and Control*, vol. 66, p. 102415, 2021.

[6] B. Karan, S. S. Sahu, and K. Mahto, "Parkinson disease prediction using intrinsic mode function based features from speech signal," *Biocybernetics and Biomedical Engineering*, vol. 40, no. 1, pp. 249–264, Jan. 2020.

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

[8] A. Hernandez, E. J. Yeo, S. Kim, and M. Chung, "Dysarthria detection and severity assessment using rhythm-based metrics," in *Proc. 21st Annual Conference of the International Speech Communication Association*, Shanghai, China, Oct. 2020, pp. 3403–3407.

[9] P. Janbakhshi, I. Kodrasi, and H. Bourlard, "Subspace-based learning for automatic dysarthric speech detection," *IEEE Signal Processing Letters*, vol. 28, no. 1, pp. 96–100, Dec. 2020.

[10] J. Vasquez, J. R. Orozco, and E. Noeth, "Convolutional neural network to model articulation impairments in patients with Parkinson's disease," in *Proc. 18th 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: Springer International Publishing, Nov. 2017, pp. 206–215.

[12] J. Mallela, A. Illa, Y. Belur, N. Atchayaram, R. Yadav, P. Reddy, D. Gope, and P. K. Ghosh, "Raw speech waveform based classification of patients with ALS, Parkinson's disease and healthy controls using CNN-BLSTM," in *Proc. 21st Annual Conference of the International Speech Communication Association*, Shanghai, China, Sep. 2020, pp. 4586–4590.

[13] P. Janbakhshi and I. Kodrasi, "Experimental investigation on STFT phase representations for deep learning-based dysarthric speech detection," in *Proc. 47th IEEE International Conference on Acoustics, Speech, and Signal Processing*, Singapore, May 2022.

[14] N. Narendra, B. Schuller, and P. Alku, "The detection of Parkinson's disease from speech using voice source information," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1925–1936, May 2021.

[15] 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, Sep. 2020.

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

[17] T. Bhattacharjee, J. Mallela, Y. Belur, N. Atchayaram, 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. 22nd Annual Conference of the International Speech Communication Association*, Brno, Czechia, Aug. 2021, pp. 2961–2965.

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

[19] P. Janbakhshi and I. Kodrasi, "Supervised speech representation learning for Parkinson's disease classification," in *Proc. 14th ITG Conference on speech communication*, Virtual Conference, Sep. 2021, pp. 1–5.

[20] P. Janbakhshi, I. Kodrasi, and H. Bourlard, "Automatic dysarthric speech detection exploiting pairwise distance-based convolutional neural networks," in *Proc. 46th IEEE International Conference on Acoustics, Speech, and Signal Processing*, Virtual Conference, May 2021, pp. 7328–7332.

[21] S. P. Dubagunta, R. J. van Son, and M. Magimai.-Doss, "Adjustable deterministic pseudonymization of speech," *Computer Speech & Language*, vol. 72, pp. 1–17, Mar. 2022.

[22] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. Patino, J. Bonastre, P. Noe, and M. Todisco, "Introducing the VoicePrivacy Initiative," in *Proc. 21st Annual Conference of the International Speech Communication Association*, Shanghai, China, Oct. 2020, pp. 1693–1697.

[23] L. Sha and T. Lukasiewicz, "Multi-type disentanglement without adversarial training," in *Proc. 35th AAAI Conference on Artificial Intelligence*, Virtual Conference, Feb. 2021, pp. 9515–9523.

[24] D. Moyer, S. Gao, R. Brekelmans, G. V. Steeg, and A. Galstyan, "Invariant representations without adversarial training," in *Proc. 32nd International Conference on Neural Information Processing Systems*, ser. NIPS'18. Red Hook, NY, USA: Curran Associates Inc., Dec. 2018, pp. 9102–9111.

[25] P. Cheng, W. Hao, S. Dai, J. Liu, Z. Gan, and L. Carin, "CLUB: A contrastive log-ratio upper bound of mutual information," in *Proc. 37th International Conference on Machine Learning*, vol. abs/2006.12013, Jul. 2020.

[26] T. yao Hu, A. Shrivastava, O. Tuzel, and C. S. Dhir, "Unsupervised style and content separation by minimizing mutual information for speech synthesis," in *Proc. 45th IEEE International Conference on Acoustics, Speech and Signal Processing*, Virtual Conference, May 2020, pp. 3267–3271.

[27] S. Yuan, P. Cheng, R. Zhang, W. Hao, Z. Gan, and L. Carin, "Improving zero-shot voice style transfer via disentangled representation learning," in *Proc. International Conference on Learning Representations*, Vienna, Austria, May 2021, pp. 1–12.

[28] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm, "Mutual information neural estimation," in *Proc. 35th International Conference on Machine Learning*, vol. 80, Jul. 2018, pp. 531–540.

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

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

[31] D. P. Kingma and J. Ba, "ADAM: a method for stochastic optimization," in *Proc. 3rd International Conference on Learning Representations*, San Diego, CA, USA, May 2015.