Subspace-based Learning for Automatic Dysarthric Speech Detection

Parvaneh Janbakhshi, *Student Member, IEEE*, Ina Kodrasi, *Member, IEEE*, and Hervé Bourlard, *Fellow, IEEE*

***Abstract*—To assist the clinical diagnosis and treatment of speech dysarthria, automatic dysarthric speech detection techniques providing reliable and cost-effective assessment are indispensable. Based on clinical evidence on spectro-temporal distortions associated with dysarthric speech, we propose to automatically discriminate between healthy and dysarthric speakers exploiting spectro-temporal subspaces of speech. Spectro-temporal subspaces are extracted using singular value decomposition, and dysarthric speech detection is achieved by applying a subspace-based discriminant analysis. Experimental results on databases of healthy and dysarthric speakers for different languages and pathologies show that the proposed subspace-based approach using temporal subspaces is more advantageous than using spectral subspaces, also outperforming several state-of-the-art automatic dysarthric speech detection techniques.**

***Index Terms*—spectral subspace, temporal subspace, Grassmann discriminant analysis, dysarthria, SVD**

I. Introduction

SPEECH dysarthria arises from disruption of muscular control needed for speech production and is a common symptom of several neurological disorders such as Parkinson's disease (PD), Amyotrophic Lateral Sclerosis (ALS), and Cerebral Palsy (CP). Depending on the origin and the severity of dysarthria, several components of the speech production mechanism such as phonation, prosody, and articulation can be affected [1]. The evaluation of these different components through auditory-perceptual assessments is used for the clinical diagnosis of dysarthria, which is crucial to adequate the management and treatment of patients. To assist the clinical diagnosis of dysarthria, automatic dysarthric speech detection techniques have gained widespread attention within the research community [2]. While most contributions deal with dysarthric speech arising due to PD [3]–[12], results on dysarthric speech arising due to ALS and CP have seldom been reported [13]–[17].

Typical automatic dysarthric speech detection techniques are based on pattern recognition methods operating on acoustic features which are hand-crafted to reflect impaired speech dimensions. Commonly used acoustic features are features characterizing impacted phonation, e.g., fundamental frequency f_0, jitter or shimmer [3]–[5], [8], [10], [15], and features characterizing impacted articulation, e.g., Mel frequency cepstral coefficients (MFCCs) [4], [5], [15]–[17]. Recently, we have proposed to jointly quantify impacted phonation and articulation by characterizing the sparsity of speech through the shape parameter of the distribution of speech spectral coefficients [18], [19]. Aiming to capture as many impaired dimensions as possible, also large-scale feature sets such as openSMILE have been used [7]–[9]. Although promising results have been reported, several issues arise in state-of-the-art automatic dysarthric speech detection techniques. Extracting features characterizing impacted phonation (e.g., f_0, jitter, or shimmer) requires voiced speech segmentation which might be unrobust due to the low quality of dysarthric speech [20], [21]. In addition, techniques which rely on a large number of acoustic features such as the openSMILE feature set suffer from an increased risk of over-fitting due to the scarcity of dysarthric speech training data.

Because of atypical changes in spectro-temporal fluctuations associated with imprecise and reduced articulatory movements in dysarthria, the dominant spectro-temporal patterns of healthy and dysarthric speech can be expected to differ [22]. Motivated by this knowledge, in this paper we propose to extract spectro-temporal subspaces spanning the dominant spectro-temporal patterns of speech and use them as acoustic features for automatic dysarthric speech detection. Using the singular value decomposition (SVD), spectro-temporal subspaces are constructed by extracting dominant basis vectors spanning the column (i.e., spectral) and row (i.e., temporal) space of the time-frequency (TF) representation. Since utterances from different speakers are unaligned and of different length, we propose to use dynamic time warping (DTW) [23] for time-alignment prior to constructing the temporal subspaces. Differently from the aforementioned state-of-the-art acoustic features, spectro-temporal subspaces can be directly extracted from continuous speech without requiring voiced speech segmentation. Further, a subspace-based representation can be robust to noise and can show better generalization performance without requiring a large amount of training data [24]–[26].

Unlike typically used acoustic features that lie in a Euclidean space, subspaces lie in a non-Euclidean space called the Grassmann manifold. Since utterances from healthy and dysarthric speakers are being represented by points on the Grassmann manifold (i.e., subspaces), classification should also be performed on this manifold to consider the structural information embedded in the subspaces. To this end, we propose to use Grassmann discriminant analysis (GDA) [27] on spectro-temporal subspaces to automatically discriminate between dysarthric and healthy speakers.

To the best of our knowledge, a subspace-based learning framework for dysarthric speech detection has never been

P. Janbakhshi and H. Bourlard are with the Idiap Research Institute, Martigny 1920, Switzerland and Ecole Polytechnique Fédérale de Lausanne, Lausanne 1015, Switzerland (e-mail: parvaneh.janbakhshi@idiap.ch, herve.bourlard@idiap.ch). I. Kodrasi is with the Idiap Research Institute, Martigny 1920, Switzerland (email: ina.kodrasi@idiap.ch). This research was supported by the Swiss National Science Foundation project no CRSII5_173711 "MoSpeeDi" on "*Motor Speech Disorders: characterizing phonetic speech planning and motor speech programming/execution and their impairments*".

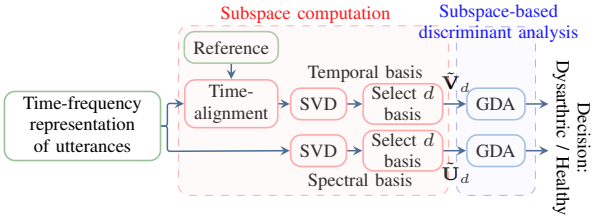

Fig. 1: Block diagram of the proposed subspace-based approach for dysarthric speech detection.

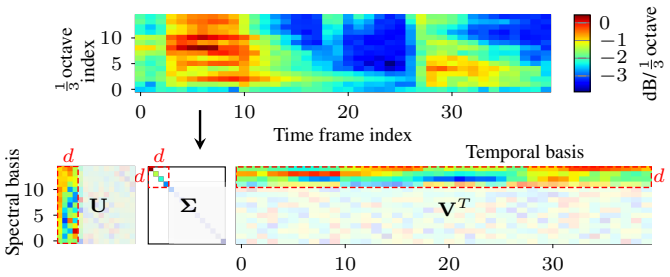

Fig. 2: Illustration of SVD for obtaining spectral and temporal basis vectors spanning the spectral and temporal dimension of the TF representation of an utterance

considered in the literature. Furthermore, while spectral subspaces are the typical choice for speech subspace analyses in many applications, temporal subspace analysis has never been explored. In [28], it has been experimentally shown that the mean of the first and second dominant spectral basis vector of healthy and dysarthric speech differ. In [29], [30], we have shown that the distance between spectral subspaces spanning pathological and intelligible speech is highly correlated with pathological speech intelligibility. However, no techniques aiming at automatic dysarthric speech detection using these spectral subspaces have been proposed.

Experimental results show that compared to spectral subspaces, temporal subspaces are more powerful discriminators for dysarthric speech detection, yielding a high performance regardless of the language or pathology and outperforming using a support vector machine (SVM) with state-of-the-art features.

II. Subspace-Based Dysarthric Speech Detection

As depicted in the schematic representation in Fig. 1, the proposed subspace-based dysarthric speech detection approach consists of computing spectro-temporal subspaces and applying subspace-based discriminant analysis using GDA. In the remainder of this section, the computational details of the proposed approach are presented.

A. Computing spectro-temporal subspaces

To obtain a signal representation resembling the transform properties of the auditory system, speech signals are first transformed to the TF domain by taking the logarithm of the one-third octave band spectrum as in [29], [30]. Let $\mathbf{S}_m$ denote the $(J \times N_m)$–dimensional TF representation of an utterance from speaker m, with J being the total number of one-third octave bands, N_m being the total number of time frames, and $J \ll N_m$. While several techniques can be used to compute spectro-temporal basis vectors, in this paper we propose to use SVD which provides an analytical solution and results in a high performance for our application. A schematic representation of applying the SVD to a sample utterance representation to obtain spectral and temporal basis vectors is depicted in Fig. 2.

The SVD of $\mathbf{S}_m$ is defined as

$$\mathbf{S}_m = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \tag{1}$$

with $\mathbf{U}$ being the $(J \times J)$–dimensional orthonormal matrix of left singular vectors, $\mathbf{\Sigma}$ being the $(J \times J)$–dimensional diagonal matrix of singular values assumed to be sorted in descending order, and $\mathbf{V}$ being the $(N_m \times J)$–dimensional orthonormal matrix of right singular vectors. Columns of $\mathbf{U}$ span the column space of $\mathbf{S}_m$, i.e., spectral space, and rows of $\mathbf{V}^T$ span the row space of $\mathbf{S}_m$, i.e., temporal space [31]. Hence, in the following, columns of $\mathbf{U}$ and $\mathbf{V}$ will be referred to as spectral and temporal basis vectors, respectively.

Spectral subspaces. To construct the spectral subspace for speaker m, $\mathbf{S}_m$ is mean-centered in each frequency band prior to computing the SVD in (1). The $(J \times d)$–dimensional matrix $\tilde{\mathbf{U}}_d$ of dominant spectral basis vectors spanning the spectral subspace is then constructed from the first d spectral basis vectors in $\mathbf{U}$, where $d < J$ since rank$(\mathbf{S}_m) = J$. The parameter d can be automatically computed based on nested cross-validation (cf. Section III-C).

Temporal subspaces. Since utterances from speakers have the same phonetic content, the dominant temporal basis vectors in $\mathbf{V}$ from (1) can be used to construct the temporal subspace from $\mathbf{S}_m$. However, temporal basis vectors obtained from different speakers cannot be directly compared to each other because of unaligned TF representations (due to different speakers and speaking rates). Therefore, prior to computing the temporal basis vectors, we propose to time-align all TF representations using DTW [23]. Following a similar procedure as in [19] for time-alignment, utterances $\mathbf{S}_m, m = 1, \ldots, M$, from all M available speakers are individually time-aligned to the $(J \times N_r)$–dimensional representation $\mathbf{S}_r$ of the same utterance from an (arbitrarily selected) healthy reference speaker r. For each time frame i in $\mathbf{S}_r$, with $i \in 1, ..., N_r$, all time frames in $\mathbf{S}_m$ that are mapped to it by DTW are extracted and averaged to create the corresponding time frame i in the time-aligned representation $\hat{\mathbf{S}}_m$. By repeating this procedure for all available $\mathbf{S}_m$, $m \neq r$, the utterance representations of all speakers are time-aligned. The dimension of the time-aligned representations $\hat{\mathbf{S}}_m$ is $J \times N_r$, i.e., it is dictated by the dimension of the reference representation $\mathbf{S}_r$.

To construct the temporal subspace for speaker m, the SVD is applied to the time-aligned representation as in (1), i.e.,

$$\hat{\mathbf{S}}_m = \hat{\mathbf{U}}\hat{\mathbf{\Sigma}}\hat{\mathbf{V}}^T, \tag{2}$$

with $\hat{\mathbf{U}}$ being a $(J \times J)$–dimensional orthonormal matrix

of spectral basis vectors, $\hat{\mathbf{\Sigma}}$ being the $(J \times J)$–dimensional diagonal matrix of singular values assumed to be sorted in descending order, and $\hat{\mathbf{V}}$ being the $(N_r \times J)$–dimensional orthonormal matrix of temporal basis vectors. Similarly to before, the time-aligned representations are mean-centered in each time frame prior to computing the SVD. The $(N_r \times d)$–dimensional matrix of dominant temporal basis vectors $\hat{\mathbf{V}}_d$ spanning the temporal subspace is then constructed from the first d temporal basis vectors in $\hat{\mathbf{V}}$, where $d < J$ since rank$(\hat{\mathbf{S}}_m) = J$. The parameter d can be automatically computed based on nested cross-validation (cf. Section III-C).

It should be noted that the computation of temporal subspaces relies on being able to accurately time-align representations. Based on our informal analyses, we are convinced that using DTW yields good alignment performance for our application.

B. Subspace-based discriminant analysis

Since subspaces lie in the Grassmann manifold which does not obey Euclidean geometry, we propose to use GDA for automatic dysarthric speech detection [27]. GDA, which has shown promising results for image classification tasks, applies kernel linear discriminant analysis (LDA) using a Grassmann kernel respecting the geometry of subspaces on the manifold. The Grassmann manifold is first mapped into a high-dimensional Hilbert space $\mathcal{H}$ which obeys Euclidean geometry. This embedded manifold is then mapped into a lower-dimensional and more discriminative Euclidean space under the Fisher LDA criteria. Finally, the dimensionality-reduced data can be classified through classical classifiers such as LDA or k-nearest neighbors [27].

For dysarthric speech detection, we are dealing with a two-class (healthy vs. dysarthric) classification problem where each class c, $c \in \{1, 2\}$, has M_c training samples (speakers). Let $\mathbf{Y}_q$ denote the orthonormal matrix representing the (spectral or temporal) subspace associated with training sample q. Further, let Φ denote the function mapping subspaces to the Hilbert space $\mathcal{H}$. Finding the discriminant Fisher direction $\mathbf{w}$ in $\mathcal{H}$ requires maximizing

$$J = \frac{\mathbf{w}^T \mathbf{S}_b^\phi \mathbf{w}}{\mathbf{w}^T \mathbf{S}_w^\phi \mathbf{w}}, \tag{3}$$

with

$$\mathbf{S}_b^\phi = \frac{1}{M} \sum_{c=1}^{2} M_c (\mathbf{m}_c^\phi - \mathbf{m}^\phi)(\mathbf{m}_c^\phi - \mathbf{m}^\phi)^T, \tag{4}$$

$$\mathbf{S}_w^\phi = \frac{1}{M} \sum_{c=1}^{2} \sum_{\mathbf{Y}_q \in c} (\mathbf{Y}_q^\phi - \mathbf{m}_c^\phi)(\mathbf{Y}_q^\phi - \mathbf{m}_c^\phi)^T, \tag{5}$$

where $M = M_1 + M_2$, $\mathbf{m}_c^\phi$ denotes the mean of the mapped training samples from class c, $\mathbf{m}^\phi$ denotes the mean of all mapped training samples, and $\mathbf{Y}_q^\phi$ denotes the mapped training sample $\mathbf{Y}_q$. Clearly, with $\mathcal{H}$ being a very high-dimensional space, (3) cannot be solved directly. To overcome this limitation, the kernel trick is used where the original subspaces $\mathbf{Y}_q$ are never explicitly mapped to $\mathcal{H}$ [32]. Instead, they are represented through a set of pairwise similarity comparisons based on a valid kernel function defined on the Grassmann manifold. The Grassmann kernel used in this paper is defined as [27]

$$k(\mathbf{Y}_p, \mathbf{Y}_q) = \left\| \mathbf{Y}_p^T \mathbf{Y}_q \right\|_F^2, \tag{6}$$

with $\{\cdot\}_F$ denoting the matrix Frobenius norm and $\mathbf{Y}_p$ and $\mathbf{Y}_q$ being the orthonormal matrices representing the (spectral or temporal) subspaces of samples p and q. Using the Grassmann kernel in (6), (3) can be reformulated without explicitly computing $\mathbf{S}_b^\phi$ and $\mathbf{S}_w^\phi$ and the discriminant direction $\mathbf{w}$ can be analytically computed and used to project the spectro-temporal subspaces onto a lower dimensional Euclidean space.[1] The final classification results presented in Section III-E are then obtained using LDA on these dimensionality-reduced subspaces.

III. Experimental Results

In this section, the performance of the proposed subspace-based approach for dysarthric speech detection is investigated and compared to state-of-the-art approaches.

A. Databases

To investigate the applicability and generalisability of the proposed approach to different pathologies and languages, the following three databases are considered.

PC-GITA database [33]. We consider Spanish recordings from 45 PD patients (22 males, 23 females) and 45 healthy speakers (22 males, 23 females) from the PC-GITA database [33]. Each speaker utters 6 sentences which are recorded at a sampling frequency of 44.1 kHz. After downsampling to 16 kHz, all sentences are concatenated and used to extract spectro-temporal subspaces and state-of-the-art features for each speaker (cf. Section III-D).

MoSpeeDi database. We consider French recordings from 20 PD and ALS patients (14 males, 6 females) and 20 healthy speakers (10 males, 10 females) from Geneva University Hospitals and University of Geneva. Each speaker utters 6 sentences which are recorded at a sampling frequency of 44.1 kHz. After downsampling to 16 kHz and manually removing non-speech segments at the beginning and end of each sentence, all sentences are concatenated and used to extract spectro-temporal subspaces and state-of-the-art features for each speaker (cf. Section III-D).

Universal access speech (UA-Speech) database [34]. We consider English recordings from 15 CP patients (11 males, 4 females) and 12 healthy speakers (8 males, 4 females) from the UA-Speech database [34]. Signals are recorded by a 7-channel microphone array at a sampling frequency of 16 kHz. For the results presented in this paper, we consider the 5th (arbitrarily selected) channel recordings of 24 words uttered by all speakers. After extracting speech-only segments using an energy-based voice activity detection [35], all words are concatenated and used to extract spectro-temporal subspaces and state-of-the-art features for each speaker (cf. Section III-D).

[1] For details on reformulating (3) using the kernel trick and computing the discriminant direction $\mathbf{w}$, the reader is referred to [32].

B. Reference speakers for time-alignment

As described in Section II-A, computing temporal subspaces requires a reference speaker for time-alignment. To avoid introducing any bias, considered reference speakers are not included in the training/testing sets of the databases described in Section III-A. To analyze the sensitivity of the temporal subspace-based approach to the reference speaker selection, 5 and 10 randomly selected (healthy) reference speakers are considered for the PC-GITA and MoSpeeDi databases, respectively. The performance of the proposed approach using each reference speaker is computed, and the presented performance values in Section III-E for the temporal subspace-based approach represent the mean and standard deviation of this performance across different reference speakers. Since a small number of speakers is available in the UA-Speech database, only 1 arbitrarily selected (healthy) reference speaker is used for the results presented on this database in Section III-E.

C. Algorithmic settings and evaluation

Spectro-temporal subspaces are extracted on the logarithm of one-third octave band representations computed using $J = 15$ and a 32 ms Hamming window with a 50% overlap (cf. II).

The validation strategy on the PC-GITA and MoSpeeDi databases is a stratified 9-fold and 4-fold cross-validation strategy, respectively. Given the small number of speakers in the UA-Speech database, the validation for this database is based on a leave-one-speaker-out strategy. The performance is evaluated in terms of the mean of accuracy, i.e., the percentage of correctly classified speakers, across all test folds.

As in [27], a regularization parameter δ is also used for GDA to avoid numerical issues and improve generalisability. Therefore, our subspace-based approach has two hyperparameters, i.e., δ and the number of basis vectors d. To select δ and d, a grid-search with $\delta \in \{10^{-10}, \ldots, 10^{-1}\}$ and $d \in \{1, \ldots, J\}$ is performed using nested cross-validation in each training fold. The final δ and d are selected as the ones yielding the highest mean accuracy on the training set.

D. State-of-the-art features

The proposed subspace-based approach is compared to using an SVM with a radial basis kernel function with state-of-the-art features such as MFCCs and the frequency-dependent shape parameter μ. When using MFCCs, the feature vector is a 56–dimensional vector constructed by extracting 4 functionals, i.e., mean, standard deviation, kurtosis, and skewness of 14 MFCCs across time [36]. When using the shape parameter μ, the feature vector is a 385–dimensional vector constructed as in [19]. For both considered feature vectors, to select the soft margin constant C and the kernel width γ of the SVM, a grid search with $C \in \{10^{-2}, \ldots, 10^{4}\}$ and $\gamma \in \{10^{-4}, \ldots, 10^{2}\}$ is performed using nested cross-validation in each training fold. The final C and γ are selected as the ones yielding the highest mean accuracy on the training set.

E. Results

Table I presents the accuracy of the considered dysarthric speech detection approaches on all considered databases, with bold entries indicating the maximum performance. The proposed spectral and temporal subspace-based approaches are denoted by S-GDA and T-GDA, respectively. For the proposed temporal subspace-based approach on the PC-GITA and MoSpeeDi databases, the mean and standard deviation of the performance across different reference speakers are also presented (cf. Section III-B). Several observations can be made based on the presented results.

TABLE I: Accuracy [%] of the proposed and state-of-the-art dysarthric speech detection methods on different databases.

Method	PC-GITA	MoSpeeDi	UA-Speech
T-GDA	$\mathbf{82.0} \pm 3.5$	$\mathbf{80.5} \pm 4.7$	$\mathbf{96.3}$
S-GDA	61.1	75.0	85.2
SVM using MFCCs	75.6	55.5	92.6
SVM using μ	72.2	67.5	88.9

First, it can be observed that the proposed subspace-based approach using temporal subspaces yields a better performance than using spectral subspaces for all considered databases. Hence, it can be said that the characterization of temporal patterns has a higher discriminative power for subspace-based healthy and dysarthric speech discrimination than the characterization of spectral patterns. Further, observing the low standard deviation of the performance of the temporal subspace-based approach suggests that this approach is not highly sensitive to the reference speaker selection. Finally, the proposed temporal subspace-based method compared to the state-of-the-art methods achieves a much better performance on all considered databases.

In summary, the presented experimental results demonstrate the applicability and advantages of the proposed subspace-based approach, with temporal subspaces yielding a better performance than spectral subspaces and also outperforming using an SVM with state-of-the-art acoustic features.

IV. Conclusion

To automatically discriminate between dysarthric and healthy speech, we have proposed a subspace-based approach representing speakers through spectral or temporal subspaces spanned by the dominant spectral or temporal basis vectors of the octave band representation of speech. Prior to constructing the temporal subspaces, it has been proposed to time-align signals to a reference representation using DTW. The spectral and temporal basis vectors are extracted using SVD. Since speakers are represented through subspaces, it has been proposed to apply subspace-based discriminant analysis to automatically discriminate between dysarthric and healthy speakers. Extensive experimental results on three databases have shown that compared to spectral subspaces, temporal subspaces are more successful in characterizing dysarthric speech. In addition, it has been shown that the proposed subspace-based approach using temporal subspaces outperforms using an SVM with state-of-the-art features for dysarthric speech detection.

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] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 5, pp. 1264–1271, May 2012.

[4] 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.

[5] D. Hemmerling, J. R. Orozco-Arroyave, A. Skalski, J. Gajda, and E. Noeth, "Automatic detection of Parkinson's disease based on modulated vowels," in *Proc. 17th Annual Conference of the International Speech Communication Association*, San Francisco, USA, Sep. 2016, pp. 1190–1194.

[6] J. R. Orozco-Arroyave, F. Hönig, J. D. Arias-Londoño, J. F. Vargas-Bonilla, and E. Noeth, "Spectral and cepstral analyses for Parkinson's disease detection in Spanish vowels and words," *Expert Systems*, vol. 32, no. 6, pp. 688–697, Mar. 2015.

[7] T. Bocklet, S. Steidl, E. Noeth, and S. Skodda, "Automatic evaluation of Parkinson's speech–acoustic, prosodic and voice related cues," in *Proc. 14th Annual Conference of the International Speech Communication Association*, Lyon, France, Aug. 2013, pp. 1149–1153.

[8] L. Berus, S. Klancnik, M. Brezocnik, and M. Ficko, "Classifying Parkinson's disease based on acoustic measures using artificial neural networks," *Sensors (Basel)*, vol. 19, no. 1, Dec. 2018.

[9] E. Vaiciukynas, A. Verikas, A. Gelzinis, and M. Bacauskiene, "Detecting Parkinson's disease from sustained phonation and speech signals," *PLOS ONE*, vol. 12, no. 10, pp. 1–16, Oct. 2017.

[10] 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.

[11] M. Novotný, J. Rusz, R. Čmejla, and E. Růžička, "Automatic evaluation of articulatory disorders in Parkinson's disease," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 9, pp. 1366–1378, Sep. 2014.

[12] F. López-Pabón, T. Arias-Vergara, and J. R. Orozco-Arroyave, "Cepstral analysis and Hilbert-Huang transform for automatic detection of Parkinson's disease," *TecnoLógicas*, vol. 23, no. 47, pp. 93–108, Jan. 2020.

[13] S. Jothilakshmi, "Automatic system to detect the type of voice pathology," *Applied Soft Computing*, vol. 21, pp. 244–249, Aug. 2014.

[14] R. Norel, M. Pietrowicz, C. Agurto, S. Rishoni, and G. Cecchi, "Detection of Amyotrophic Lateral Sclerosis (ALS) via acoustic analysis," in *Proc. 19th Annual Conference of the International Speech Communication Association*, Hyderabad, India, Sep. 2018, pp. 377–381.

[15] J. Wang, P. Kothalkar, B. Cao, and D. Heitzman, "Towards automatic detection of Amyotrophic Lateral Sclerosis from speech acoustic and articulatory samples," in *Proc. 17th Annual Conference of the International Speech Communication Association*, San Francisco, USA, Sep. 2016, pp. 1195–1199.

[16] A. Illa, D. Patel, B. K. Yamini, S. S. Meera, N. Shivashankar, P.-K. Veeramani, S. vengalii, K. Polavarapui, S. Nashi, A. Nalini, and P. K. Ghosh, "Comparison of speech tasks for automatic classification of patients with Amyotrophic Lateral Sclerosis and healthy subjects," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Calgary, Canada, Apr. 2018, pp. 6014–6018.

[17] S. Gillespie, Y.-Y. Logan, E. Moore, J. Laures-Gore, S. Russell, and R. Patel, "Cross-database models for the classification of dysarthria presence," in *Proc. 18th Annual Conference of the International Speech Communication Association*, Stockholm, Sweden, Aug. 2017, pp. 3127–3131.

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

[19] ——, "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.

[20] V. Parsa and D. G. Jamieson, "Acoustic discrimination of pathological voice: sustained vowels versus continuous speech." *Journal of Speech, Language, and Hearing Research*, vol. 44, no. 2, pp. 327–339, Apr. 2001.

[21] J. R. Orozco-Arroyave, J. F. Vargas-Bonilla, and E. Delgado-Trejos, "Acoustic analysis and non linear dynamics applied to voice pathology detection: A review," *Recent Patents on Signal Processing*, vol. 2, no. 12, pp. 96–107, Sep. 2012.

[22] K. M. Rosen, R. D. Kent, A. L. Delaney, and J. R. Duffy, "Parametric quantitative acoustic analysis of conversation produced by speakers with dysarthria and healthy speakers," *Journal of Speech Language and Hearing Research*, vol. 49, no. 2, pp. 395–411, Apr. 2006.

[23] L. Rabiner and B. Juang, *Fundamentals of speech recognition*. New Jersely, USA: Prentice-Hall, Aug. 1993.

[24] Ruiping Wang, Shiguang Shan, Xilin Chen, and Wen Gao, "Manifold-manifold distance with application to face recognition based on image set," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, June 2008, pp. 1–8.

[25] S. Chen, C. Sanderson, M. T. Harandi, and B. C. Lovell, "Improved image set classification via joint sparse approximated nearest subspaces," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Portland, OR, USA, June 2013, pp. 452–459.

[26] B. Mishra, H. Kasai, P. Jawanpuria, and A. Saroop, "A riemannian gossip approach to subspace learning on grassmann manifold," *Machine Learning*, vol. 108, no. 10, pp. 1783–1803, Jan. 2019.

[27] J. Hamm and D. D. Lee, "Grassmann discriminant analysis: A unifying view on subspace-based learning," in *Proc. 25th International Conference on Machine Learning*, Helsinki, Finland, July 2008, pp. 376–383.

[28] A. Kacha, F. Grenez, J. R. Orozco-Arroyave, and J. Schoentgen, "Principal component analysis of the spectrogram of the speech signal: Interpretation and application to dysarthric speech," *Computer Speech & Language*, vol. 59, pp. 114–122, Jan. 2020.

[29] P. Janbakhshi, I. Kodrasi, and H. Bourlard, "Spectral subspace analysis for automatic assessment of pathological speech intelligibility," in *Proc. 20th Annual Conference of the International Speech Communication Association*, Graz, Austria, Sep. 2019, pp. 3038–3042.

[30] ——, "Automatic pathological speech intelligibility assessment exploiting subspace-based analyses," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, no. 1, pp. 1717–1728, May 2020.

[31] A. J. Van Der Veen, E. F. Deprettere, and A. L. Swindlehurst, "Subspace-based signal analysis using singular value decomposition," *Proceedings of the IEEE*, vol. 81, no. 9, pp. 1277–1308, Sep. 1993.

[32] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K. R. Mullers, "Fisher discriminant analysis with kernels," in *Proc. 1999 IEEE Signal Processing Society Workshop*, Madison, WI, USA, Aug. 1999, pp. 41–48.

[33] 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.

[34] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," in *Proc. 9th Annual Conference of the International Speech Communication Association*, Brisbane, Australia, Sep. 2008, pp. 1741–1744.

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

[36] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in OpenSMILE, the Munich Open-Source Multimedia Feature Extractor," in *Proc. 21st ACM International Conference on Multimedia*, Barcelona, Spain, Oct. 2013, pp. 835–838.