Multiview Canonical Correlation Analysis for Automatic Pathological Speech Detection

Yacouba Kaloga, Shakeel A. Sheikh, Ina Kodrasi
Idiap Research Institute, Martigny, Switzerland
{yacouba.kaloga, shakeel.sheikh, ina.kodrasi}@idiap.ch

***Abstract*—Recently proposed automatic pathological speech detection approaches rely on spectrogram input representations or wav2vec2 embeddings. These representations may contain pathology-irrelevant uncorrelated information, such as changing phonetic content or variations in speaking style across time, which can adversely affect classification performance. To address this issue, we propose to use Multiview Canonical Correlation Analysis (MCCA) on these input representations prior to automatic pathological speech detection. Our results demonstrate that unlike other dimensionality reduction techniques, the use of MCCA leads to a considerable improvement in pathological speech detection performance by eliminating uncorrelated information present in the input representations. Employing MCCA with traditional classifiers yields a comparable or higher performance than using sophisticated architectures, while preserving the representation structure and providing interpretability.**

***Index Terms*—pathological speech detection, interpretability, MCCA, Parkinson's disease.**

I. Introduction

Speech production is a complex process where the brain executes a series of sequences involving various sensorimotor, muscle, and articulator processes. The execution of these sequences can be affected by neurodegenerative disorders such as Parkinson's disease (PD) [1], which can result in pathological speech characterized by imprecise articulation, insufficient prosody, and other abnormal speech patterns [2], [3]. In clinical settings, diagnosis of neurological impairments is typically done through auditory-perceptual tests alongside other meta data such as genetic information and visual cues [4]. However, diagnostic accuracy can vary among clinicians depending on their experience, implicit biases, as well as their condition during the diagnosis [5]. To address this issue, researchers are exploring various automatic pathological speech detection approaches. These approaches mainly differ in the input representations and the classifers that are exploited.

In traditional automatic pathological speech detection approaches based on machine learning, handcrafted acoustic features often inspired by clinical knowledge are fed to classical algorithms, such as e.g., support vector machines (SVMs) [6], multi layer perceptrons (MLPs) [7], or random forests [8]. Various feature sets have been examined, including openSMILE [9], Mel-frequency cepstral coefficients [10], or sparsity-based features [6]. Despite the reported promising performance, hand-crafted features have a limited capability to comprehensively capture nuances, cues, and complexities of pathological speech.

Deep learning (DL)-based approaches on the other hand have the potential to capture more abstract and subtle cues of pathological speech. These approaches commonly employ spectrogram input representations such as the short-time Fourier transform (STFT) or Mel spectrograms [6], [11]. Spectrograms capture both temporal and spectral information, making them well-suited for analyzing and interpreting speech signals [12]. Approaches exploiting spectrogram input representations often rely on complex architectures [8]. For example, convolutional neural networks (CNNs) have been exploited in [13], [14] to leverage the two-dimensional structure of the time-frequency representations. In [15], long short-term memory (LSTMs) networks are used for their capability to capture long-range dependencies. Despite the higher performance reported by these approaches compared to approaches using handcrafted features, they no longer represent the state-of-the-art in the field. Today, approaches relying on self-supervised learning (SSL) models such as wav2vec2 (w2v2) are preferred [16]–[18]. These models leverage a vast collection of available audio data to learn embeddings which enable unprecedented performance for several downstream tasks [19]. Despite their performance, interpreting SSL features remains challenging, discouraging their deployment in clinical practice. Consequently, approaches based on spectrogram input representations remain relevant and efforts are still directed at improving their performance.

Speech representations such as spectrograms and w2v2 embeddings include cues about speaker identity, phonetic content, and emotional state. However, many of these cues can be irrelevant or even detrimental for pathological speech detection. To mitigate this issue, supervised adversarial and non-adversarial training methods have been proposed to suppress these speaker identity cues [20], [21], while dimensionality reduction techniques such as principal component analyses (PCA) and linear discriminant analysis have been used to remove other redundant information [22], [23]. Although dimensionality reduction techniques are well-studied for handcrafted acoustic features, they remain under-explored for DL-based input representations. This is because DL-based approaches are expected to be powerful enough to ignore pathology-unrelated cues in input representations. However, due to the limited data typically available for this task, effectively learning to do so remains challenging.

In this paper, we propose to use multiview canonical correlation analysis (MCCA) [24] to remove redundant information from input speech representations prior to training pathological speech detection approaches. While MCCA has been extensively investigated in the context of computer vision [25], recommendation systems [26], silent speech recognition [27] and various classification tasks [28], to the best of our knowledge, it has never been incorporated in pathological speech detection. MCCA, a multiview version of PCA, finds a common representation across multiple views, which is often more effective for classification or clustering than using any single view or their concatenation [29]. We hypothesize that unlike pathology-discriminant cues, pathology-irrelevant cues such as changing phonetic content or speaking style are uncorrelated over time. By considering different chunks of an utterance as separate views of the speaker within the MCCA framework, MCCA representations preserve the correlated pathology-discriminant cues while suppressing the uncorrelated pathology-irrelevant cues. This approach

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

maintains the structure of the representation, enabling state-of-the-art performance with simpler models, like MLPs, without losing interpretability.

II. Multiview Canonical Correlation Analysis

Elements of a dataset can be described from various "viewpoints", such as images of the same object taken from different angles or the perception of an event acquired with two or more sensory inputs. The representations of the same element from different perspectives can be considered as distinct views. The objective of MCCA is to find a shared low-dimensional representation from distinct views of the element.

Given two views $X_1 \in \mathbb{R}^{f\times t_1}$ and $X_2 \in \mathbb{R}^{f\times t_2}$, we aim to find the optimal projectors $U_1 \in \mathbb{R}^{t_1\times t}$ and $U_2 \in \mathbb{R}^{t_2\times t}$, where $t \leq \min(t_1, t_2)$, such that the correlation between X_1U_1 and X_2U_2 is maximized. This objective can be expressed as the optimization problem

$$\min_{U_1,U_2} \|X_1U_1 - X_2U_2\|_2^2 \text{ s.t. } U_m^T(X_m^TX_m)U_m = I_t, \qquad (1)$$

with $m \in \{1, 2\}$ and I_t being the $t \times t$–dimensional identity matrix. The solution to (1) is obtained through an eigendecomposition, as described in [29]. Extending this approach to multiview data presents challenges, as maximizing the pairwise correlation between $M > 2$ different views is NP-hard [30]. The MAXVAR relaxation of this problem [24] seeks a shared low-dimensional representation $S \in \mathbb{R}^{f\times t}$ that closely approximates each low-dimensional projection X_mU_m, where $m = 1, \ldots, M$. This leads to MCCA, which is formulated as

$$\min_{S,(U_m)_{m=1\ldots M}} \sum_{m=1}^{M} \|X_mU_m - S\|_2^2 \text{ s.t } S^TS = I_t. \qquad (2)$$

The optimal representation S^* solving (2) is obtained by extracting the columns corresponding to the m-leading eigenvectors of the matrix $\sum_{m=1}^{M} X_m^T(X_mX_m^T)^{-1}X_m$ as described in [29][1], whereas the optimal projection matrices U_m^* are computed as $U_m^* = (X_mX_m^T)^{-1}X_mS^{*T}$. As shown in [28], [29], using the low-dimensional representation S^* instead of the individual views X_m (or their concatenation) yields an advantageous performance.

III. Multiview Canonical Correlation Analysis for Pathological Speech Detection

We consider a neurotypical and pathological speech dataset containing n elements, denoted as $\{\boldsymbol{x}_i\}_{i=1}^n$. Each element $\mathbf{x}_i \in \mathbb{R}^{F\times T_i}$ represents a segment of speech data as e.g., a spectrogram or w2v2 embedding, where F indicates the feature dimensionality and T_i the number of time frames. Each representation is labeled as neurotypical or pathological. The number of time frames T_i can be different among the different elements of the dataset. However, in order to increase the number of elements in the dataset, we consider representations of fixed-size segments of speech extracted from full utterances as in [10], [20]. State-of-the-art literature directly uses these representations to train pathological speech detection models. However, these representations contain extensive pathology-irrelevant cues related to the varying phonetic content, speaking style, or emotional state of the speaker, which can be detrimental to pathological speech detection performance. Our assumption in this paper is that some irrelevant cues for pathological speech detection do not exhibit temporal self-correlation. Hence, to remove uncorrelated irrelevant cues from these representations, we propose to incorporate MCCA as described in the following.

[1]Note that $X_mX_m^T$, the covariance matrix of view m, is typically regularized as $X_mX_m^T + 10^{-4}I_{t_m}$ for numerical stability.

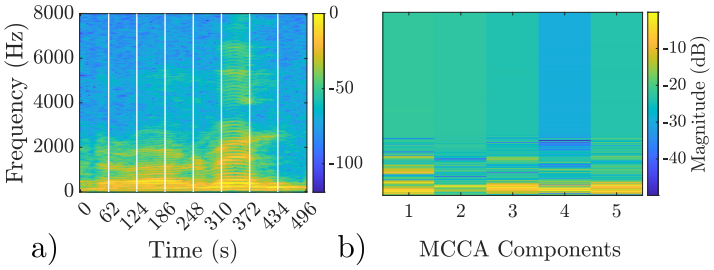

Fig. 1: (a) Spectrogram of a 500 ms long speech segment with 257 frequency bins representing the frequency range from 0 Hz to 8000 Hz. The vertical lines denote the boundary of $M = 8$ chunks, with each chunk considered to be a single view of the speech segment. (b) Representation after applying MCCA, corresponding to the same 257 bins and containing $T = 5$ components.

To use MCCA for pathological speech detection, we divide the $F \times T_i$–dimensional representation $\boldsymbol{x}_i$ into M chunks $\boldsymbol{x}_i^{(m)} \in \mathbb{R}^{F\times\lfloor\frac{T_i}{M}\rfloor}$, with M being a user-defined parameter (cf. Section V-A), $m \in 1, \ldots, M$, and $\lfloor\cdot\rfloor$ denoting the floor operator. If M does not divide T_i, the remaining time frames are discarded. We consider these chunks to be the M distinct views (denoted by X_m in Section II) in the context of MCCA. After applying MCCA (cf. (2)), the original $F \times T_i$–dimensional representation $\mathbf{x}_i$ is reduced to the $F \times T$–dimensional representation $\boldsymbol{x}_i^*$ (denoted by S^* in Section II), with T representing the number of MCCA components. These dimensionality-reduced representations of speech segments can then be used as input to classification models such as MLP or LGBM. The choice of the number of chunks M reflects the time span along which relevant cues are expected to be correlated while irrelevant cues are expected to be uncorrelated. By carefully selecting M, we can eliminate irrelevant cues from input representations and improve the performance of pathological speech detection approaches. However, different representations exhibit different characteristics and correlation spans, making it challenging to derive a general optimal number of chunks M from a theoretical perspective. Experimental results in Section V-A provide insights into this matter.

Fig. 1a depicts an exemplary spectrogram of a speech segment (i.e., the representation used by state-of-the-art approaches based on the STFT). For a user-defined number of chunks $M = 8$, the vertical lines depicted in Fig. 1a show the boundaries between the different chunks that are considered to be the different views of this spectrogram. Applying MCCA with $T = 5$ components to these chunks yields the representation in Fig. 1b. As expected, MCCA preserves the feature structure (i.e., along the F dimension) and contains most of its variance in the low frequency components. Since MCCA preserves the feature structure, it enables the identification of features that influence the decision of classifiers. This is particularly important when using spectrogram representations, where the F features represent the energy in different frequency bins.

IV. Experimental Settings

In the following, we present the various experimental settings used for generating the experimental results in Section V.

Dataset. We use Spanish recordings from a group of 50 patients diagnosed with Parkinson's disease (25 males, 25 females) along with 50 neurotypical speakers (25 males, 25 females) from the PC-GITA database [31]. Spontaneous speech recordings of speakers

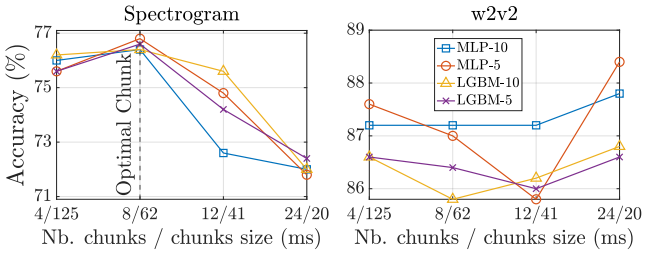

Fig. 2: MLP and LGBM performance for different chunk sizes M with 5 and 10 MCCA components using (a) spectrogram and (b) w2v2 embeddings.

discussing their day are used. Recordings are downsampled to 16 kHz and segmented into 500 ms segments with a 50% overlap prior to computing input representations. The average length of the total available speech material for each speaker is 47.1 s.

Input representations and detection models. We investigate the applicability of the proposed MCCA approach to both spectrogram and w2v2 input representations. For the spectrogram representations, the STFT is computed using a 32 ms Hamming window with a hop size of 4 ms, resulting in 257×126–dimensional representations. For the w2v2 representations, we extract embeddings from the first layer of the transformer module of the XLRS53 version of w2v2 as in [18], resulting in 1024×24–dimensional representations. These input representations (and their MCCA versions) are used to train two different pathological speech detection models; an MLP, which is a common choice in pathological speech research [10], and LGBM, which is known for its superior performance with careful parameter selection [32].

Evaluation. Evaluation is performed using speaker-independent stratified 10-fold cross-validation, with one fold as the test set, one as the validation set, and the remaining eight as the training set. This experiment is repeated with 5 different seeds and we report the average and standard deviation of the performance across these different seeds. The performance is evaluated in terms of speaker-level accuracy, which is computed through soft voting of the probability of decisions for all segments belonging to each speaker.

Training. For each input representation and detection model, we conduct hyperparameter tuning on the validation set (as specified below), with specific grid searches for each model focusing on key parameters to balance training time and performance.

- MLP: Hyperparameters include the number of hidden layers $\{2, 3, 4\}$ with $\{64, 128\}$ units per layer. The maximum number of iterations is fixed at 3000.
- LGBM: Hyperparameters include *num_leaves* $= 31$, *min_child_samples* $\in \{2, 3\}$, *max_depth* $\in \{20, 30\}$, *n_estimators* $\in \{400, 500, 600\}$, *colsample_bytree* $\in \{0.1, 0.2\}$, *learning_rate* $= 0.01$, *is_unbalanced* $=$ *True*, and *boosting type* $=$ *Dart*.

All remaining MLP and LGBM hyperparameters are set to default values from their respective *scikit-learn* 1.2.2 and *LightGBM* 4.2.0 libraries.

TABLE I: MLP and LGBM performance using PCA and MCCA with different number of components on spectrogram representations.

		Components				
Models	**Features**	1	2	3	5	10
MLP	PCA	64.40 ±3.44	67.20 ±3.43	68.20 ±1.17	69.20 ±2.71	71.00 ±1.79
	MCCA	68.20 ±1.60	74.40 ±2.50	76.20 ±3.43	76.8 ±2.32	76.40 ±3.00
LGBM	PCA	68.60 ±1.02	70.20 ±1.60	70.20 ±2.14	70.60 ±2.50	70.60 ±2.65
	MCCA	72.20 ±1.94	75.80 ±1.33	75.60 ±0.49	76.60 ±1.62	76.40 ±1.20

V. Experimental Results and Discussion

In the following, the pathological speech detection performance achieved when applying MCCA to input representations is extensively investigated. In Section V-A we analyze the impact of the number of chunks M for computing the MCCA representations on the detection performance. In Section V-B the detection performance achieved using MCCA is compared to the traditionally used PCA. In Section V-C the performance using MCCA is further improved through feature selection and insights on interpretability are provided.

A. Impact of the number of chunks

As previously mentioned, to apply MCCA to input representations, we first need to determine the optimal number of chunks M, which corresponds to the time scale over which correlation is maximized. To this end, we evaluate the performance of MLP and LGBM classifiers using spectrogram and w2v2 input representations for different chunk sizes, i.e., $M = 4$ (125 ms), $M = 8$ (62 ms), $M = 12$ (41 ms), and $M = 24$ (20 ms). To ensure that the derived optimal number of chunks for each representation are not specific to a single number of MCCA components, we also consider two different numbers of components, i.e., 5 and 10. The number of chunks, along with the other hyperparameters described in Section IV, are optimized on the validation set. Such a procedure results in $M = 8$ as the optimal M for spectrogram representations and $M = 24$ as the optimal M for w2v2 representations, regardless of the classifier used and number of MCCA components. Fig. 2 shows the performance on the test set for each M for all considered input representations, classification models, and number of MCCA components, where it is confirmed that $M = 8$ and $M = 24$ yield the best performance for spectrogram and w2v2 represetations, respectively.

This difference in the optimal M value for the two representations can be attributed to their different temporal resolution, with a spectrogram frame representing no contextual information and a w2v2 embedding representing contextual information. As a result, it is expected that correlation should be maximized on a larger time scale for the spectrogram representation than for the w2v2 representation. In the remainder of this section, we report only results using these optimal chunk sizes computed on the validation set (i.e., $M = 8$ for spectrogram and $M = 24$ for w2v2 representations).

TABLE II: MLP and LGBM performance using PCA and MCCA with different number of components on w2v2 representations.

		Components				
Models	**Features**	1	2	3	5	10
MLP	PCA	83.20 ±1.94	84.60 ±1.62	84.60 ±2.24	85.60 ±2.24	87.20 ±1.17
	CCA	82.80 ±2.04	84.40 ±1.83	86.60 ±1.62	88.40 ±1.55	87.80 ±1.47
LGBM	PCA	86.00 ±1.41	85.40 ±0.49	85.20 ±1.17	85.80 ±1.17	85.60 ±1.72
	CCA	86.60 ±1.74	86.40 ±1.67	86.60 ±0.49	86.60 ±0.49	86.80 ±1.72

B. MCCA vs PCA

In this section we validate the applicability of MCCA in comparison to the traditional PCA dimensionality reduction for pathological speech detection. The performance when using MLP and LGBM with different number of PCA or MCCA components on spectrogram and w2v2 representations are presented in Tables I and II.

Table I shows that for spectrogram input representations, using MCCA considerably outperforms using PCA, independently of the number of components or classifier used. These results validate our hypothesis that certain pathology-irrelevant information in spectrogram input representations is temporally uncorrelated, with MCCA suppressing it, and hence, improving the performance of pathological speech detection. It is worth mentioning that using spectrogram input representations with more complex architectures like CNN yields a speaker-level accuracy of 69.72% on the PC-GITA database with the same experimental settings as ours [11]. Hence, these results also show that using MCCA and simple classifiers such as MLP or LGBM on spectogram input representations results in a considerable performance improvement (of 6% to 7%) in comparison to using more complex architectures like CNN.

Table II shows that for w2v2 representations, using MCCA yields a similar or a slightly better performance than using PCA, independently of the number of components or classifier used. These results reinforce our hypothesis that employing MCCA helps in selecting more relevant information across time, even when using powerful multilingual SSL embeddings which already yield an impressive performance.

C. Important Features and Interpretation

Since MCCA is beneficial for pathological speech detection as shown in Section V-B, in this section we further improve its performance and explore its interpretability. For the analysis presented in this section, we focus on using LGBM and MCCA (with 5 components) on spectrogram and w2v2 representations.

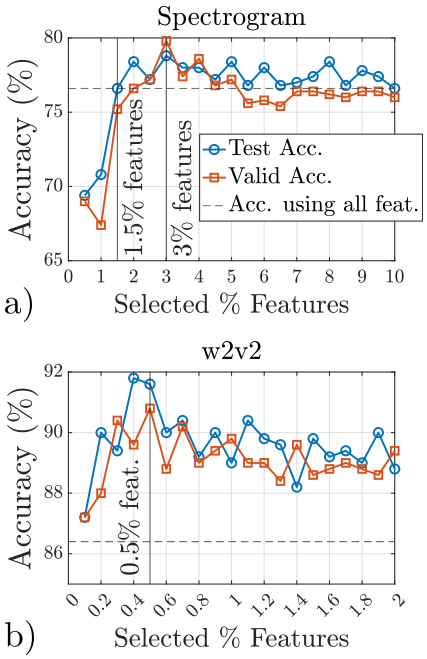

Fig. 3: LGBM performance on the validation and test sets using different % of top-ranked features for (a) spectrogram and (b) w2v2 embeddings. For ease of comparison, the performance on the test set using all features is also illustrated.

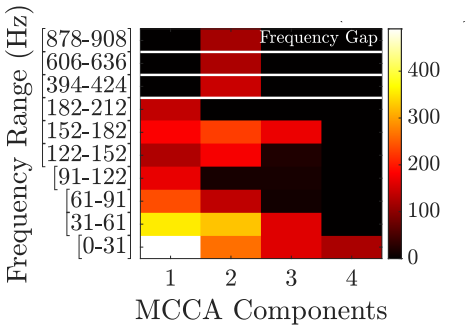

Fig. 4: MCCA components of the frequency bins belonging to the 1.5% top-ranked features. The color map illustrates the importance assigned to each bin when using LGBM.

Using each of the 50 models (10 folds $\times$ 5 seeds) trained in our previous experiments, we calculate the average importance assigned by LGBM to each of the features across the different training sets. From this ranking, we train new models using only a percentage k of the top-ranked features. For spectrogram representations, we consider $k \in [0.5, 1, \ldots, 10]\%$, while for w2v2 representations, we consider $k \in [0.1, 0.2, \ldots, 2]\%$. The optimal hyperparameters found in the previous experiment are reused to avoid repeating the grid search.

Fig. 3 depicts the performance (on the validation and test sets) obtained when using only the k top-ranked features as well as the test set performance when using all features (previously reported). For spectrogram input representations, Fig. 3a shows that with only 1.5% of the features (i.e., 18 features vs. 32382 in the spectrogram and 1285 after MCCA), we achieve test performance levels similar to using all features. These 18 top features are shown in Fig. 4, which depicts the MCCA components only for the frequency bins belonging to the 1.5% top-ranked features. It can be observed that these features are located in the bottom-left corner of the MCCA representation, with frequencies ranging from 0 to 210 Hz, corresponding to the average frequency of human voice, being highly important as expected. This plot shows that the choice of MCCA is relevant, as the most correlated features, i.e., component 1 and 2, are found to be the most important. For w2v2 input representations, Fig. 3b shows that using any of the k top-ranked features improves the test performance in comparison to using all features. More importantly, Fig. 3 shows that selecting the percentage of features that yields the best validation performance, i.e., 3% (56) for spectrogram and 0.5% (256) for w2v2 embeddings, results in an improved test set accuracy, reaching 78.8% and 91.6% respectively. These values surpass those obtained in our previous experiments, and notably, exceed what has been reported with such approaches in the literature to the best of our knowledge. These findings underscore MCCA's effectiveness in extracting meaningful features for PD detection using simple classifiers while preserving the essential feature structure, which is vital for further analysis. Future work will explore the use of the more powerful non-linear MCCA methods proposed in [33].

In this paper we have proposed to incorporate MCCA in state-of-the-art pathological speech detection approaches based on spectrogram and w2v2 input representations. The presented results show that MCCA improves the performance through preserving pathology-discriminant cues and discarding pathology-irrelevant information that is uncorrelated across time. More powerful MCCA methods and generalization and robustness in noisy conditions remain topics to investigate in the future.

REFERENCES

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

[2] K. Tjaden, "Speech and swallowing in Parkinson's disease," *Topics in Geriatric Rehabilitation*, vol. 24, no. 2, pp. 115–126, Nov. 2008.

[3] Y. Yunusova *et al.*, "Articulatory movements during vowels in speakers with dysarthria and healthy controls," *Journal of Speech and Hearing Research*, vol. 51, no. 3, pp. 596–611, June 2008.

[4] G. T. Stebbins and C. G. Goetz, "Factor structure of the unified Parkinson's disease rating scale: motor examination section," *Movement Disorders: Journal of the Movement Disorder Society*, vol. 13, no. 4, pp. 633–636, July 1998.

[5] M. Pernon, F. Assal, I. Kodrasi, and M. Laganaro, "Perceptual classification of motor speech disorders: the role of severity, speech task, and listener's expertise," *Journal of Speech, Language, and Hearing Research*, vol. 65, no. 8, pp. 2727–2747, 2022.

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

[7] A. Farhadipour, H. Veisi, M. Asgari, and M. A. Keyvanrad, "Dysarthric speaker identification with different degrees of dysarthria severity using deep belief networks," *Journal of Electronics and Telecommunications Research Institute*, vol. 40, no. 5, pp. 643–652, July 2018.

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

[9] N. N P and P. Alku, "Glottal source information for pathological voice detection," *IEEE Access*, vol. 8, pp. 67 745–67 755, April 2020.

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

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

[12] X. Huang, A. Acero, H.-W. Hon, and R. Reddy, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, 1st ed. USA: Prentice Hall PTR, 2001.

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

[14] B. F. Zaidi, S. A. Selouani, M. Boudraa, and M. Sidi Yakoub, "Deep neural network architectures for dysarthric speech analysis and recognition," *Neural Computing and Applications*, vol. 33, pp. 9089–9108, Aug. 2021.

[15] U. K. Lilhore, S. Dalal, N. Faujdar, M. Margala, P. Chakrabarti, T. Chakrabarti, S. Simaiya, P. Kumar, P. Thangaraju, and H. Velmurugan, "Hybrid CNN-LSTM model with efficient hyperparameter tuning for prediction of Parkinson's disease," *Scientific Reports*, vol. 13, p. 14605, Sept. 2023.

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

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

[18] S. A. Sheikh and I. Kodrasi, "Impact of speech mode in automatic pathological speech detection," in *Proc. European Signal Processing Conference*, Lyon, France, Aug. 2024, pp. 3127–3131.

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

[20] P. Janbakhshi and I. Kodrasi, "Supervised speech representation learning for Parkinson's disease classification," in *Proc. Speech Communication; 14th ITG Conference*. VDE, 2021.

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

[22] 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.

[23] M. K. Arjmandi *et al.*, "An optimum algorithm in pathological voice quality assessment using wavelet-packet-based features, linear discriminant analysis and support vector machine," *Biomedical Signal Processing and Control*, vol. 7, no. 1, pp. 3–19, Jan. 2012.

[24] J. R. Kettenring, "Canonical analysis of several sets of variables," *Biometrika*, vol. 58, no. 3, pp. 433–451, Dec. 1971.

[25] X. Yan, S. Hu, Y. Mao, Y. Ye, and H. Yu, "Deep multi-view learning methods: A review," *Neurocomputing*, vol. 448, pp. 106–129, Aug. 2021.

[26] J. Chen, G. Wang, and G. B. Giannakis, "Graph multiview canonical correlation analysis," *IEEE Transactions on Signal Processing*, vol. 67, no. 11, pp. 2826–2838, Jun. 2019.

[27] M. Kim, B. Cao, T. Mau, and J. Wang, "Multiview representation learning via deep cca for silent speech recognition," in *Proc. Annual Conference of the International Speech Communication*, 2017, pp. 2769–2773.

[28] Y. Kaloga, P. Borgnat, S. P. Chepuri, P. Abry, and A. Habrard, "Variational graph autoencoders for multiview canonical correlation analysis," vol. 188, p. 108182, Nov. 2021.

[29] J. Chen, G. Wang, and G. B. Giannakis, "Graph multiview canonical correlation analysis," *IEEE Transactions on Signal Processing*, vol. 67, no. 11, pp. 2826–2838, June 2019.

[30] J. Rupnik, P. Skraba, J. Shawe-Taylor, and S. Guettes, "A comparison of relaxations of multiset cannonical correlation analysis and applications," *arXiv preprint arXiv:1302.0974*, 2013.

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

[32] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: a highly efficient gradient boosting decision tree," in *Proc. Annual Conference on Neural Information Processing Systems*, Red Hook, NY, USA, Dec. 2017, p. 3149–3157.

[33] Y. Kaloga, P. Borgnat, S. P. Chepuri, P. Abry, and A. Habrard, "Variational graph autoencoders for multiview canonical correlation analysis," *Signal Processing*, vol. 188, p. 108182, Nov. 2021.