Detection of Neurogenic Voice Disorders Using the Fisher Vector Representation of Cepstral Features

[†,*]Madhu Keerthana Yagnavajjula, [*]Paavo Alku, [‡]Krothapalli Sreenivasa Rao, and [‡]Pabitra Mitra, *Espoo, Finland, and †‡Kharagpur, India

Summary: Neurogenic voice disorders (NVDs) are caused by damage or malfunction of the central or peripheral nervous system that controls vocal fold movement. In this paper, we investigate the potential of the Fisher vector (FV) encoding in automatic detection of people with NVDs. FVs are used to convert features from frame level (local descriptors) to utterance level (global descriptors). At the frame level, we extract two popular cepstral representations, namely, Mel-frequency cepstral coefficients (MFCCs) and perceptual linear prediction cepstral coefficients (PLPCCs), from acoustic voice signals. In addition, the MFCC features are also extracted from every frame of the glottal source signal computed using a glottal inverse filtering (GIF) technique. The global descriptors derived from the local descriptors are used to train a support vector machine (SVM) classifier. Experiments are conducted using voice signals from 80 healthy speakers and 80 patients with NVDs (40 with spasmodic dysphonia (SD) and 40 with recurrent laryngeal nerve palsy (RLNP)) taken from the Saarbruecken voice disorder (SVD) database. The overall results indicate that the use of the FV encoding leads to better identification of people with NVDs, compared to the defacto temporal encoding. Furthermore, the SVM trained using the combination of FVs derived from the cepstral and glottal features provides the overall best detection performance.

Key Words: Neurogenic voice disorders−MFCC−Perceptual linear prediction−Fisher vector−Support vector machine−Glottal features.

INTRODUCTION

Neurogenic voice disorders (NVDs) are organic voice disorders that result from problems with the central or peripheral nervous system innervation to the larynx therefore affecting functioning of the vocal mechanism.[1,2] NVDs can be the only or first sign indicating that a person has a neurological condition.[3,4] Examples of neurological conditions that show voice disorders during the disease progression are vocal tremor, spasmodic dysphonia (SD), and vocal fold paralysis.[1,2] In patients with NVDs, typical vocal signs and symptoms include quality issues such as hoarseness and harshness; vocal effort issues such as vocal fatigue and breathy voice; and pitch issues such as pitch breaks, and inappropriately high pitch[1−4] that all greatly affect the patient's ability to communicate. The assessment of voice is essential in distinguishing speakers with NVDs from healthy speakers. The voice assessment can be performed using a classical approach, which involves an otolaryngologist performing intelligibility tests to assess for abnormalities in articulation, quality, rate of speech and fluidity (whether there are breaks or spasms).[4] Subjective intelligibility tests are, however, costly, laborious, and frequently prone to intrinsic biases of physicians due to familiarity with patients and their voice condition. This motivates the design of machine learning (ML)-based systems that could automatically detect individuals with NVDs directly from acoustic voice signals thereby enabling objective assessment. Objective voice-based assessment is economical and reliable, and it can be used to perform the diagnosis away from the hospital, which reduces the inconvenience and cost of frequent physical visits of patients for medical examination.[4] Most importantly, ML-driven automated NVD detection systems can be used to screen individuals with neurogenic disorders at an early stage, which helps in providing timely treatment for the patients.

Typically, ML-based automatic detection of voice disorders includes two major components[5,7]: feature extraction and classifier. In feature extraction, a set of pre-defined features are extracted to capture discriminative information present in voice signals. The feature sets reported in the literature for detection of voice disorders can be grouped into four categories: (1) spectral and cepstral measures (such as Mel-frequency cepstral coefficients (MFCCs), linear predictive cepstral coefficients (LPCCs), cepstral peak prominence (CPP) and perceptual linear prediction cepstral coefficients (PLPCCs));[5−8] (2) perturbation measures (such as the jitter and shimmer);[11−13] (3) complexity measures (such as the Hurst exponent, approximate entropy, and sample entropy);[14,15] and (4) glottal source measures (such as time-domain and frequency-domain glottal source parameters).[5,7,9,10] Among various features, the cepstral features (particularly MFCCs) are most popular and they have been shown to perform comparably to or better than many other feature types.[16,17] The cepstral domain representations have the advantage that they can effectively capture the abnormalities in articulation and vocal quality (such as irregular vocal fold movements and incomplete closure of the vocal

Accepted for publication October 21, 2022.

From the *Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland; †Advanced Technology Development Centre, Indian Institute of Technology Kharagpur, Kharagpur, West Bengal, India; and the ‡Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur, Kharagpur, West Bengal, India.

Address correspondence and reprint requests to Madhu Keerthana Yagnavajjula. E-mail addresses: madhu.yagnavajjula@aalto.fi, madhu.keerthu@iitkgp.ac.in

Journal of Voice, Vol. ■■, No. ■■, pp. ■■−■■
0892-1997

https://doi.org/10.1016/j.jvoice.2022.10.016

folds[6]). Furthermore, the cepstral features are less correlated, which is advantageous in the efficient implementation of ML classifiers. Therefore, in this study we make use of the cepstral features in distinguishing healthy speakers from patients with NVDs. The classifier stage includes an ML algorithm trained with the extracted features to label the input voice as healthy or disordered. For the classifier stage, several algorithms such as support vector machine (SVM), artificial neural networks, decision trees, and variants of recurrent neural network (RNN) have been used in the study area.[5,6,8,9,18−20] Among various classifiers, SVM is the most widely used algorithm for detection of voice disorders.[5,10,18,19] A review by Al-Dhief et al.[18] provides more information of various feature extraction and classification techniques that have been used in detection of pathological voice.

The utterance-level features (also known as global descriptors) derived from the frame-level features (also known as local descriptors),[5,10,17] such as MFCCs, are commonly used for voice disorder detection tasks. Global descriptors can embed long-range dependencies present in voice signals, and capture the most relevant information from the entire utterance in a compact form. Furthermore, global descriptors can present variable-length voice signals using a fixed-length vector, which enables easy training of ML classifiers. Most of the existing studies utilize temporal encoding that uses descriptive statistics (such as mean and standard deviation) to convert frame-level features to utterance-level features. The main disadvantage of temporal encoding is that it mixes the information from the local descriptors and cannot capture their joint variation with time. It is important to note that the information specific to a voice disorder is not equally distributed among all frame-level features extracted from a voice sample. For example, in a single voice sample of a patient with NVD there may be some feature frames with no dysphonia and some with varying degrees of dysphonia. Hence, it is necessary to have a representation which preserves this temporal difference without blending the information. In this work, we propose to use the *Fisher vector* (FV) *encoding* for deriving the global descriptors.[21] Unlike temporal encoding, FV encoding can capture finer relationship between each feature dimension along with temporal pattern.[22] Further, FV is efficient to compute and can provide accurate encoding even with a small corpus. By utilizing FV representation, best results have been reported in para-linguistics tasks such as speaker verification[24] and spoofing detection.[22] Since detection of NVD is also a para-linguistic task,[23] we hypothesize that better discrimination of healthy speakers from patients with NVD can be achieved by using the FV encoding technique.

The paper is organized as follows. Section 2 describes the database used for the detection task. Section 3 provides the details about the NVD detection system, the considered local descriptors, the two encoding techniques (temporal and FV) for converting local descriptors to global descriptors, SVM classification, and evaluation criteria. The results are reported in Section 4. Finally, the conclusions of this study are provided in Section 5.

DATABASE

In our experiments, we considered the publicly available Saarbruecken voice disorder (SVD) database,[25] which in total consists of recordings from 869 healthy and 1356 pathological speakers of German. The database is a large repository of pathological speech containing recordings of the sustained vowels /a/, /i/, and /u/ in high, normal, and low pitches, as well as with rising-falling pitch. Furthermore, the database also comprises recordings of the sentence ”Guten Morgen, wie geht es Ihnen?” (”Good morning, how are you?”). Speech signals representing as many as 71 different pathologies are present in the SVD database. For the current study, we used a portion of the database by selecting voices produced by healthy speakers and by patients suffering from NVDs. The number of speakers considered in the study is 80 for healthy speakers (40 male and 40 female) and 80 for speakers with NVD (40 male and 40 female). From the 80 speakers diagnosed with NVD, 40 suffer from spasmodic dysphonia (a central nervous system disorder) and 40 suffer from recurrent laryngeal nerve palsy (a peripheral nervous system disorder). SD causes involuntary spasms in the muscles of the voice box or larynx. This causes the voice to break and have a tight, strained or strangled sound.[4] Vocal fold paralysis due to dysfunction of recurrent laryngeal nerve is referred to as recurrent laryngeal nerve palsy/paralysis (RLNP).[4] Patients with RLNP typically complain of hoarseness, changes in vocal pitch, and breathy quality of voice. The speaker age ranges from 30 to 80 yr (mean 56.03 and 55.39 for the healthy speakers and the speakers with NVD, respectively; standard deviation 14.51 and 15.19 for the healthy speakers and the speakers with NVD, respectively). The number of male and female speakers is same in both groups (healthy and NVD). For each speaker, the current study utilizes the three vowels produced with normal, low and high pitch and the sentence yielding in total 10 utterances per speaker (nine vowel utterances and one sentence). A sampling frequency of 50 kHz was used in the original recordings of the SVD database, but we down-sampled all the signals to 16 kHz for the purpose of this study.

ARCHITECTURE OF THE NVD DETECTION SYSTEM

Figure 1 depicts the steps in the system for the automatic detection of NVDs. In the training phase, local descriptors (frame-level features) are first extracted from voice signals. As local descriptors, we considered 13-dimensional MFCCs and 13-dimensional PLPCCs extracted from the acoustic voice signal, and 13-dimensional MFCCs extracted from the glottal source signal estimated by means of glottal inverse filtering.[26] The extraction process of the local descriptors is discussed in Section 3.1. The local descriptors are converted to global descriptors (utterance-level features) using either the temporal or FV encoding (discussed in Section 3.2). Finally, an SVM classifier is trained using individual or combined global descriptors along with the corresponding class labels (healthy vs. NVD). In the testing phase, the global descriptors of an utterance are given as

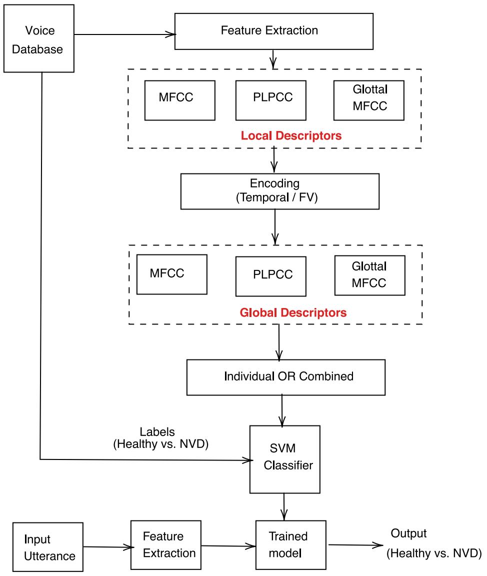

FIGURE 1. Schematic block diagram of the NVD detection system.

input to SVM, which detects if the input voice signal is produced by a healthy speaker or by a speaker with NVD.

Extraction of local descriptors

This subsection describes the extraction of local descriptors used for the detection task of the current study.

The MFCC and PLPCC features

Figure 2 shows the steps involved in extraction of the MFCC features. The input speech signal is first pre-emphasized and divided into several 30 ms frames using the Hamming window and a hop size of 10 ms. Next, the 1024-point discrete Fourier transform (DFT) of each frame is computed. After this, a triangular filter bank consisting of 40 Mel-spaced filters is applied to the power spectrum. Finally, through the discrete cosine transform (DCT) calculation of the logarithm of filterbank output, 13 MFCCs are obtained for each frame.

The computation process of 13 PLPCCs is similar to that of the MFCC computation, except that it involves two additional steps: equal-loudness pre-emphasis and intensity loudness conversion prior to applying the logarithm (see Figure 2). The perceptual enhancement (PE) block shown in Figure 2 integrates these two steps. The purpose of pre-emphasizing the spectrum is to approximate the unequal sensitivity of human hearing to different frequencies. The intensity-to-loudness conversion is intended for tuning of the spectral envelope approximation. Both MFCCs and PLPCCs have been shown to provide very good discrimination of healthy and disordered voices,[16] and are regarded as the defacto standard feature sets in the study area. Figure 3 (a) shows the spectrogram computed from voice signals of a healthy speaker and a patient with NVD. From Figure 3(a), it can be seen that the spectrograms of the healthy and disordered voices are clearly different. The MFCC and PLPCC try to capture the differences in the spectra in their own way.

The glottal-MFCC features

Recent studies have shown that the glottal source signal carries complementary information of voice disorders.[5,7,10] In this work, we capture this information through MFCCs computed from the glottal source waveform derived using glottal inverse filtering (GIF). For GIF, we considered the quasi-closed phase (QCP) technique which has been shown to compute glottal source signals from non-modals voices better than several existing techniques.[9,26] For the details of the QCP method, the reader is referred to the study by Airaksinen et al.[26] As seen from Figure 2, the steps involved in the computation of the 13-dimensional glottal-MFCCs (shortly referred to as gMFCCs) is the same as that of the MFCC computation from acoustic voice signals, except that the input signal is the glottal source signal instead of the voice signal. Figure 3(c) shows spectrograms of the glottal source signals estimated using the QCP method from the healthy and disordered voice signals shown in Figure 3(b). Like in the spectrograms computed from voice signals shown in Figure 3(a), the spectra of the glottal source signals show clear differences. The discriminatory information present in the spectrum of the glottal source signal is represented by the glottal-MFCC features.

Deriving global descriptors from local descriptors

The MFCC, PLP and glottal-MFCC features (described in previous sections) are called local descriptors as they capture frame-level information. For the detection of voice disorders from longer utterances like sustained vowels, words and sentences, utterance-level information is desired. Therefore, local descriptors are converted to global descriptors,

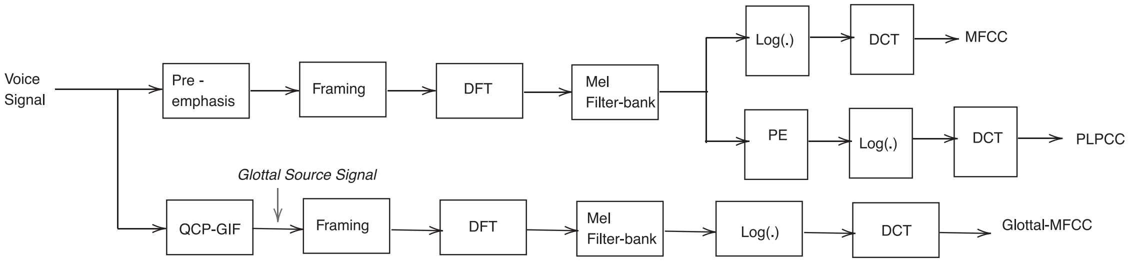

FIGURE 2. Block diagram representation for the extraction of different cepstral features used as local descriptors in the study. The PE and log(.) blocks denote perceptual enhancement and logarithm operation, respectively.

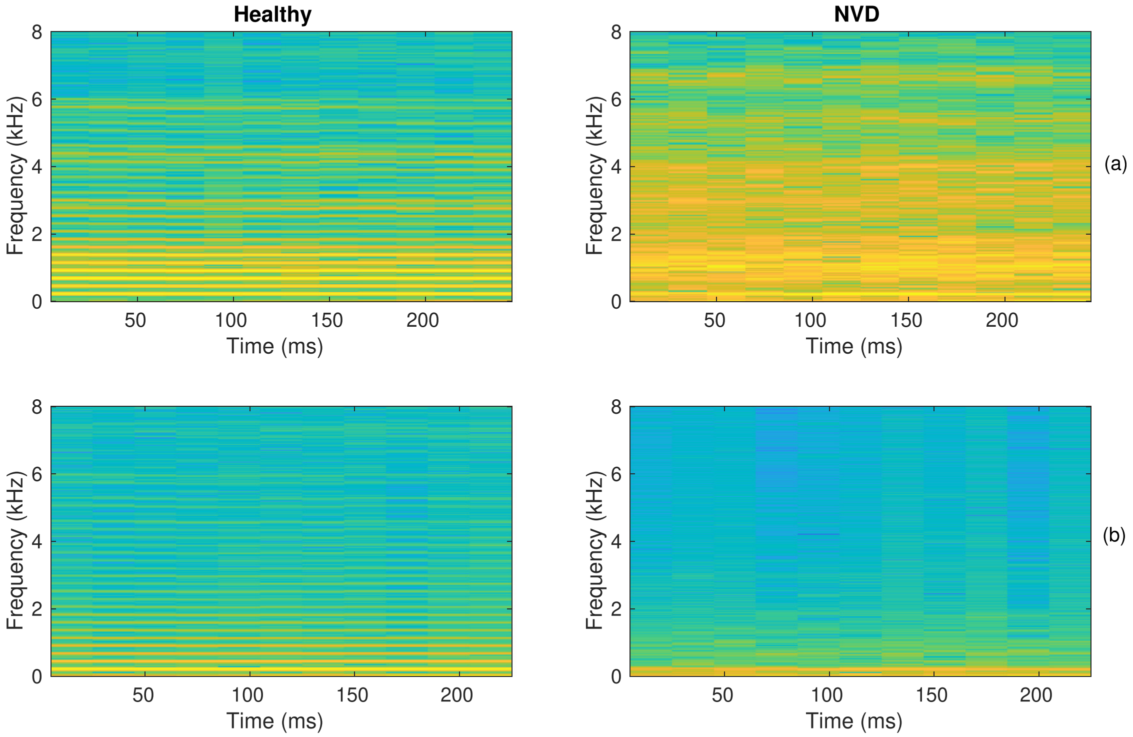

FIGURE 3. Illustration of spectrograms of (a) acoustic voice signals of vowel /a/ and (b) their corresponding glottal source waveforms estimated using the QCP method, for a healthy speaker and a patient with NVD caused by RLNP.

which condense the frame-level representations into a single utterance level representation. The global descriptors helps in better characterization of healthy/disordered voices compared to local descriptors by representing the long-term trends in the local descriptors over an utterance. Conventionally, temporal encoding is used for converting local descriptors to global descriptors. In the temporal encoding, multiple descriptive statistics are considered in order to capture different aspects of voice signals.[5,10,17] For experiments, we considered four popular statistical measures, namely, mean, standard deviation, skewness, and kurtosis. The length of a global descriptor is equal to DN where D is the dimension of the feature vector representing each frame and N is the number of statistical measures used. For the considered local descriptors, each frame consists of a 13-dimensional feature vector, so the global descriptors have a dimension of 52. That is, each input voice signal is represented using a 52-dimensional feature vector irrespective of the length of the input signal.

The main drawback of the temporal encoding is that it models the temporal trajectories of each dimension in the feature vectors independently. As a result, the temporal encoding approach fails to capture the joint variation of the cepstral coefficients with time. As an alternative to the temporal encoding, we make use of the FV encoding technique for deriving global descriptors. In the case of the FV encoding, all the dimensions of the feature vectors are considered together and a multivariate Gaussian is fitted. Hence, the FV encoding captures a finer relationship between each feature dimension along with temporal pattern. Also, FV encoding uses the information from a larger population of each class while fitting a Gaussian, and hence can better represent the inter-class variability than temporal encoding. The main idea behind the FV encoding is to measure the amount of change induced by the utterance descriptors on a background probability model, which is typically a Gaussian Mixture Model (GMM) with diagonal covariance matrices. The Fisher vector encodes the amount of change of the model parameters to optimally fit the new-coming data. This requires the computation of the Fisher information matrix, which is the derivative of the log-likelihood with respect to model parameters (hence the name “Fisher”). First, a GMM model with K Gaussians is learned using the training set of any one of the class (healthy or NVD). The GMM model is parameterized as $\lambda = \{w_k, \mu_k, \sigma_k\}_{k=1}^{K}$ where w_k, μ_k, σ_k represent mixture weight, mean and variance corresponding to the k-th Gaussian component, respectively. Once the model is trained, the FV representation of a set of local descriptors $L = \{l_1, l_2, ...l_F\}$, where F is the number of frames, is given by two parts:[21,22]

$$u_k = \frac{1}{N\sqrt{w_k}} \sum_{k=1}^{K} q_{ki} \left(\frac{l_i - \mu_k}{\sigma_k} \right) \tag{1}$$

$$v_k = \frac{1}{N\sqrt{2w_k}} \sum_{i=1}^{N} q_{ki} \left\{ \left(\frac{l_i - \mu_k}{\sigma_k} \right)^2 - 1 \right\} \tag{2}$$

where q_{ki} is the Gaussian soft assignment of the descriptor x_i to the kth Gaussian. The u part captures the 1st order differences whereas the v part captures the second order differences. With a D-dimensional local descriptor, the final FV representation of size $2DK$ is obtained by concatenation of the two parts. In this study, two Gaussian components ($K = 2$) are used to fit the feature vectors for all local descriptors extracted from patient's data. Hence, as in the case of the temporal encoding, a 52-dimensional feature vector is extracted from every input voice signal with the FV encoding. The FVs are standardized before utilizing for training the SVM classifier.

SVM classifier

The databases for voice disorders usually contain a small number of speech samples. SVM, a fast and reliable ML algorithm, performs very well with a limited amount of data, and hence it has become one of the most popular classifiers for detection of voice disorders. The SVM algorithm tries to find the optimal placement of the separation plane between the borders of two classes to achieve the maximum discrimination. In this work, we used a non-linear SVM with the radial basis function (RBF) kernel. The kernel equation is given by

$$K(x, y) = exp(-\gamma \parallel x - y \parallel^2 \Big), \quad \gamma > 0 \tag{3}$$

where the training samples, labels and kernel parameter are denoted by x, y, and γ respectively. Besides the kernel parameter, there is also a regularization parameter C for SVM. The SVM classifier is trained with the global descriptors computed from the voice signals as input and the corresponding class labels as output.

Evaluation scheme

The evaluation was carried out using the five-fold cross validation (CV) technique. In each fold of the CV, data from 80% of the speakers (64 speakers consisting of 32 male and 32 female speakers) was used for training and data from the remaining 20% of the speakers (16 speakers consisting of eight male and eight female speakers) from each class was used for testing. Every speaker was used only once for testing and the same speaker was not used in both training and testing. For evaluation we considered four standard metrics, namely, recall (sensitivity), precision, F1-score and accuracy. These metrics are defined as follows:

$$recall(\%) = \frac{TP}{TP + FN} \times 100 \tag{4}$$

$$precision(\%) = \frac{TP}{TP + FP} \times 100 \tag{5}$$

$$F1 - score(\%) = 2\frac{precision * REC}{precision + REC} \times 100 \tag{6}$$

$$accuracy(\%) = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \tag{7}$$

In the equations above, TP refers to true positives (ie, the number of samples which are truly positive and are predicted as positive), TN refers to true negatives (ie, the number of samples which are truly negative and are predicted as negative); FP refers to false positives (ie, the number of samples which are truly negative but are predicted as positive) and FN refers to false negatives (ie, the number of samples which are truly positive but are predicted as negative). For a good detection system, all the metrics should be high (ideally close to 100%) Evaluation metrics were saved in every fold and were averaged over the five folds for the evaluation. The hyper-parameters of the SVM classifier were tuned automatically using the Bayesian optimization technique by following a 10-fold CV strategy using the training data from the first fold.

RESULTS

The performance of the temporal encoding and FV encoding techniques is evaluated using the voice signals from the SVD database as discussed in Section 2. The average NVD detection results obtained with the individual and combined feature sets are shown in Table 1. From the table, it can be observed that in the case of individual feature sets, the PLPCC feature set provided the best performance in terms

TABLE 1.
Performance Evaluation Results (zin %) Obtained for the Detection Task (Healthy vs. NVD)

Features	Temporal Encoding				FV Encoding			
	Recall	Precision	F1-Score	Accuracy	Recall	Precision	F1-Score	Accuracy
				Individual Feature Sets				
MFCC	68.50	68.01	68.24	68.16	69.17	72.17	70.64	71.25
PLPCC	69.33	69.10	69.22	69.17	71.67	72.88	72.27	72.50
gMFCC	66.50	65.32	65.87	65.58	67.67	66.81	67.10	66.92
				Combined Feature Sets				
MFCC + PLPCC	74.67	70.60	72.50	71.58	75.17	**75.21**	75.04	75.00
MFCC + gMFCC	73.33	67.69	70.40	69.17	71.00	74.15	72.49	73.08
PLPCC + gMFCC	75.83	70.54	73.09	72.08	77.50	72.66	75.00	74.17
MFCC + PLPCC + gMFCC	75.50	71.86	73.63	72.92	**80.83**	74.05	**77.29**	**76.25**

For each of the four metrics, the best combination of encoding and features is marked in bold.

of accuracy (69.17%), F1-score (69.22%), precision (69.10%) and recall(69.33%) with the temporal encoding. In terms of the F1-score and accuracy, the next best feature set was MFCC, which provided an accuracy of 68.17% and F1-score of 68.24%. The performance of gMFCC was also close to that of the MFCC and PLPCC features, which indicates the presence of voice pathology related information in the gMFCC features. Most importantly, the results obtained for all the three individual feature sets show that the global descriptors generated by the FV encoding yielded better performance than those obtained with the temporal encoding. Among the individual feature sets, the overall best performance in terms of recall (71.67%), precision (72.88%), F1-score (72.27%), and accuracy (72.50%), was obtained by using the FV representation for the PLPCC features.

From the combination of feature sets, it can be clearly seen that there exists an improvement in performance for all the combinations with the temporal and FV encoding. This indicates existence of complementary information among the feature sets. In the case of the temporal encoding, the combination of FVs of PLPCC and gMFCC provided better performance compared to the combinations MFCC +PLPCC and MFCC+gMFCC. However, with the FV encoding, the detection performance achieved with the combination (MFCC+gMFCC) was better than the other two combinations. In the case of both the temporal and FV encodings, the best performance was observed when the MFCC, PLPCC and glottal-MFCC features were combined. This highlights the complementary nature of the conventional cepstral features with the glottal source cepstral features for the NVD detection. It can be noted that both with individual or combined feature sets the detection performance achieved with the FV encoding is better than the one achieved with the temporal encoding. Overall, combining the FV representations of all the three feature sets provided the best detection performance in terms of accuracy (76.25%) and F1-score (77.29%). Figure 4 shows the receiver operating character (ROC) curves for the FV and temporal encoding of the combination (MFCC + PLPCC + gMFCC). The ROC curves are plotted by considering data of 75% of the speakers for training and the data from remaining speakers for testing. From the figure, it can be seen that the FV encoding achieved the best area under curve (AUC) of 0.83, and the ROC curves demonstrate the superiority of the FV encoding over the temporal encoding. Altogether, the results highlight the importance of capturing the joint spectro-temporal variations between features for the detection of voice disorders.

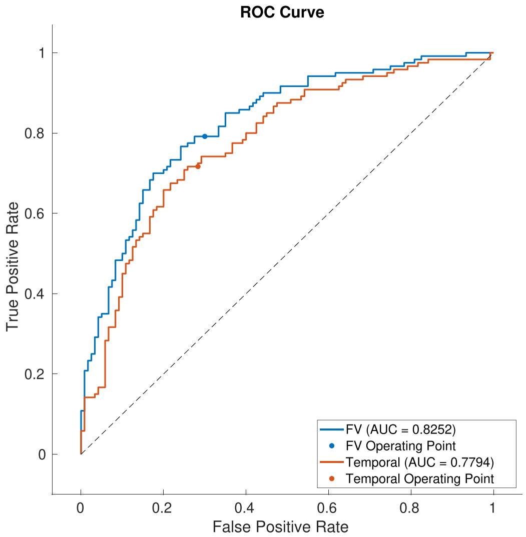

FIGURE 4. The ROC curves for classification by SVM.

DISCUSSION

Neurological disorders affect the body systemically, but patients will often complain of dysphonia before other symptoms develop. Therefore, detection of people with NVDs is essential to provide timely treatment for the underlying neurological condition. The ML-based detection of NVDs is a non-invasive approach for distinguishing speakers with NVDs from healthy controls automatically using the acoustic voice signal. This paper studies the effectiveness of cepstral features and the FV representation in the detection of NVDs. Patients with two types of NVDs, namely, SD (a central nervous system disorder) and RLNP (a peripheral nervous system disorder) are considered. While the patients with SD have a strangled, harsh-sounding voice, with inappropriate pitch or pitch breaks, those with RLNP have voice quality that is hoarse and breathy.[4] The abnormalities in the voices of patients with NVDs can be captured effectively using the cepstral features (MFCC and PLPCC), which have the ability to parameterize changes in the articulation and phonation caused by the neurologic conditions.[6]

For the voice disorder detection task, it is desirable to condense frame-level information into utterance-level information. Therefore, most of the existing studies make use of temporal encoding to convert the frame-level features to utterance-level features to capture long-term dependencies. However, the temporal encoding technique merges the important temporal cues and hence cannot efficiently represent the discriminative information in voice signals. To overcome this drawback, we used the FV encoding technique, which generates global descriptors by modelling the subtle variations of local descriptors over time. The experiments results show that the FV representation of the cepstral features results in better performance compared to temporal encoding. Since the considered NVDs (SD and RLNP) mainly affect phonation, we further examined the effectiveness of voice source information in the detection of patients with NVD. The voice source was derived using the QCP inverse filtering technique and the estimated glottal flows were parameterised using MFCC features. As the major finding of this study, the results showed that the FV representation is more effective in the detection of NVDs than the temporal encoding. Furthermore, the combination of FV representations of MFCC, PLPCC and glottal-MFCC features provided the overall best performance in terms of recall (80.83%), F1-score (77.29%), and accuracy (76.25%). It is important to note that the majority of the samples used for experiments were sustained vowels. Sustained vowel production is not necessarily the best type of speaking task to elicit voice signals that could reveal the presence of abnormal temporal behaviour caused by a NVD. Even for a clinician, it might be difficult to perceptually assess sustained vowel samples. On the other hand, sustained vowels are easy to collect from patients. In addition, sustained vowels are present in spontaneous speech in all languages. Therefore they can in principle be recorded from natural everyday speech communication situations unlike voice samples that call conducting a special speaking task such the diadochokinetic task.[9] Therefore, the studied combination of the FV representations of the MFCC, PLPCC and glottal-MFCC features can be considered beneficial because it provides an effective detection scheme that can automatically flag individuals with abnormal vocal characteristics already from simple and short sustained utterances.

CONCLUSIONS

In this study, we investigated the effectiveness of the FV encoding in developing automatic ML-based systems for the detection of people with NVDs. The FV representation is used for converting local descriptors extracted from voice signals to global descriptors, which can capture the joint spectro-temporal variations. The FV representations of the popular MFCC and PLPCC as well as the glottal-MFCC features were employed in the detection of NVDs. The experimental were performed using voice signals of healthy speakers and patients with two types of NVDs (SD and RLNP), present in the SVD database. The results indicate that the FV encoding technique performs better than the defacto temporal encoding technique. Furthermore, the combination of FVs of cepstral features or FVs of cepstral and glottal features resulted in improved detection performance, indicating the complementary nature of these features. The combination of FVs of all the considered features provided the overall best performance in discriminating healthy speakers and speakers with NVDs.

Acknowledgments

This research was supported by the Academy of Finland (project no. 330139), Aalto University (the MEC program for India) and Tata Consultancy Services India (the TCS Research Scholar Program).

REFERENCES

1. Gamboa J, Jiménez FJ, Mate MA, et al. Alteraciones de la voz causadas por enfermedades neurológicas [voice disorders caused by neurological diseases]. *Rev Neurol(Ed impr)*. 2001;10:153–168.
2. Hanson DG. Neuromuscular disorders of the larynx. *Otolaryngol Clin North Am*. 1991;24:1035–1051.
3. Barkmeier JM, Case JL, Ludlow CL. Identification of symptoms for spasmodic dysphonia and vocal tremor: a comparison of expert and nonexpert judges. *J Commun Disord*. 2001;34:21–37.
4. Wang TV, Song PC. Neurological voice disorders: a review. *Int J Head Neck Surg*. 2022;13:32–40.
5. Reddy MK, Alku PA. A comparison of cepstral features in the detection of pathological voices by varying the input and filterbank of the cepstrum computation. *IEEE Access*. 2021;9:135953–135963.
6. Reddy MK, Helkkula P, Keerthana YM, et al. The automatic detection of heart failure using speech signals. *Comput Speech Lang*. 2021;69:101205.
7. Wu Y, Zhou C, Fan Z, et al. Investigation and evaluation of glottal flow waveform for voice pathology detection. *IEEE Access*. 2021;9:30–44.
8. Fraile R, Godino-Llorente JI, Sáenz-Lechón N, et al. Spectral analysis of pathological voices: sustained vowels vs running speech. In: *Proceedings of the Seventh International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*. IEEE; 2011.
9. Narendra NP, Schuller B, Alku P. The detection of Parkinson's disease from speech using voice source information. *IEEE/ACM Trans Audio Speech Lang Process*. 2021;29:1925–1936.
10. Narendra NP, Alku P. Glottal source information for pathological voice detection. *IEEE Access*. 2020;8:67745–67755.
11. Silva DG, Oliveira LC, Andrea M. Jitter estimation algorithms for detection of pathological voices. *EURASIP J Adv Signal Process*. 2009;2009:1–9.
12. Vasilakis M, Stylianou Y. Voice pathology detection based on short term jitter estimations in running speech. *Folia Phoniatr Logop*. 2009;61:153–170.
13. Zhang Y, Jiang JJ, Biazzo L, et al. Perturbation and nonlinear dynamic analyses of voices from patients with unilateral laryngeal paralysis. *J Voice*. 2005;19:519–528.
14. Arias-Londoño JD, Godino-Llorente JI, Sáenz-Lechón N, et al. Automatic detection of pathological voices using complexity measures, noise parameters, and mel-cepstral coefficients. *IEEE Trans Biomed Eng*. 2011;58:370–379.
15. Arias-Londoño JD, Godino-Llorente JI. Entropies from markov models as complexity measures of embedded attractors. *Entropy*. 2015;17:3595–3620.
16. Gómez-García JA, Moro-Velázquez L, Godino-Llorente JI. On the design of automatic voice condition analysis systems. part II: review of speaker recognition techniques and study on the effects of different variability factors. *Biomed Signal Process Control*. 2019;48:128–143.
17. Monge-Álvarez J, Hoyos-Barceló C, Lesso P, et al. Robust detection of audio-cough events using local hu moments. *IEEE J Biomed Health Inform*. 2019;23:3595–3620.
18. Al-Dhief FT, Latiff NMAA, Malik NNNA, et al. A survey of voice pathology surveillance systems based on internet of things and machine learning algorithms. *IEEE Access*. 2020;8:64514–64533.
19. Reddy MK, Alku PA, Rao KS. Detection of specific language impairment in children using glottal source features. *IEEE Access*. 2020;8:15273–15279.
20. Mayle A, Mou Z, Bunescu RC, et al. Diagnosing dysarthria with long short-term memory networks. *Proc Interspeech*. 2019:4514–4518.
21. Perronnin F, Dance C. Fisher kernels on visual vocabularies for image categorization. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition.*20072007:1–8.
22. Alam J. On the use of fisher vector encoding for voice spoofing detection. In: *Proc. International Conference on Ubiquitous Computing and Ambient Intelligence (UCAmI 2019)*2019.
23. Schuller B, Steidl S, Batliner A, et al. Paralinguistics in speech and languagestate-of-the-art and the challenge. *Comput Speech Lang*. 2013;27:4–39.
24. Tian Y, He L, Li Z, et al. Speaker verification using fisher vector. In: *Proc. International Symposium On Chinese Spoken Language Processing.*. IEEE; 2014:419–422.
25. Pützer M, Barry WJ, Available online: u-s. Saarbrücken voice database, institute of phonetics. *Univ Saarland*. 2010.
26. Airaksinen M, Raitio T, Story B, et al. Quasi closed phase glottal inverse filtering analysis with weighted linear prediction. *IEEE/ACM Trans Audio Speech Lang Process*. 2014;22:596–607.