Mittapalle, Kiran; Alku, Paavo

A Comparison of Cepstral Features in the Detection of Pathological Voices by Varying the Input and Filterbank of the Cepstrum Computation

Published in:
IEEE Access

DOI:
10.1109/ACCESS.2021.3117665

Published: 01/01/2021

Document Version
Publisher's PDF, also known as Version of record

Please cite the original version:
Mittapalle, K., & Alku, P. (2021). A Comparison of Cepstral Features in the Detection of Pathological Voices by Varying the Input and Filterbank of the Cepstrum Computation. *IEEE Access*, *9*, 135953 - 135963. https://doi.org/10.1109/ACCESS.2021.3117665

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 21, 2021, accepted September 30, 2021, date of publication October 4, 2021, date of current version October 11, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3117665

A Comparison of Cepstral Features in the Detection of Pathological Voices by Varying the Input and Filterbank of the Cepstrum Computation

MITTAPALLE KIRAN REDDY AND PAAVO ALKU, (Fellow, IEEE)
Department of Signal Processing and Acoustics, Aalto University, 00076 Espoo, Finland

Corresponding author: Mittapalle Kiran Reddy (kiran.r.mittapalle@aalto.fi)

This work was supported by the Academy of Finland under Project 330139.

ABSTRACT Automatic voice pathology detection enables objective assessment of pathologies that affect the voice production mechanism. Detection systems have been developed using the traditional pipeline approach (consisting of the feature extraction part and the detection part) and using the modern deep learning -based end-to-end approach. Due to the lack of vast amounts of training data in the study area of pathological voice, the former approach is still a valid choice. In the existing detection systems based on the traditional pipeline approach, the mel-frequency cepstral coefficient (MFCC) features can be regarded as the *defacto* standard feature set. In this study, automatic voice pathology detection is investigated by comparing the performance of various MFCC variants derived by considering two factors: the input and the filterbank in the cepstrum computation. For the first factor, three inputs (the voice signal, the glottal source and the vocal tract) are compared. The glottal source and the vocal tract are estimated using the quasi-closed phase glottal inverse filtering method. For the second factor, the mel-frequency and linear-frequency filterbanks are compared. Experiments were conducted separately using six databases consisting of voices produced by speakers suffering from one of four disorders (dysphonia, Parkinson's disease, laryngitis, or heart failure) and by healthy speakers. Support vector machine (SVM) was used as the classifier. The results show that by combining mel- and linear-frequency cepstral coefficients derived from the glottal source and vocal tract, better overall detection accuracy was obtained compared to the *defacto* MFCC features derived from the voice signal. Furthermore, this combination provided comparable or better performance than four existing cepstral feature extraction techniques in clean and high signal-to-noise ratio (SNR) conditions.

INDEX TERMS Voice disorders, glottal inverse filtering, support vector machine, cepstral coefficients.

I. INTRODUCTION

Voice pathologies arise either due to physical changes in the voice production mechanism (e.g., in the respiratory system, vocal folds, and vocal tract) [1], [2] or due to improper vocal use when the physical structure of the mechanism is normal (e.g., vocal fatigue or ventricular phonation) [3]–[5]. Examples of voice pathologies are dysarthria [7], dysphonia [8], vocal polyp [9], and developmental dysphasia [13]. Voice pathology may also indicate early neurodegenerative disease such as Parkinson's disease (PD) [10]–[12], [14]. Voice pathology detection refers to a technology to automatically distinguish normal voices from pathological voices by computer using the recorded voice signal. Existing voice pathology detection systems can be divided into two categories: traditional pipeline systems and modern end-to-end systems [15].

The traditional pipeline system consists of two components [15], [16]: the feature extraction part and the detection part. The feature extraction part tries to capture discriminative information from acoustic voice signal waveforms by representing this information in compressed forms using a set of pre-defined features. The feature sets reported in the literature for voice pathology detection can be grouped into four categories: (1) perturbation measures (such as jitter and shimmer); (2) spectral and cepstral measures

The associate editor coordinating the review of this manuscript and approving it for publication was Shuihua Wang.

(such as mel-frequency cepstral coefficients (MFCC), linear predictive cepstral coefficients (LPCC), cepstral peak prominence (CPP) and perceptual linear prediction cepstral coefficients (PLPCC)); (3) complexity measures (such as the Hurst exponent, approximate entropy, and sample entropy); and (4) glottal source measures (such as time-domain and frequency-domain glottal source parameters) [11], [15–26]. The detection part includes a machine learning (ML) classifier to label the input voice as healthy or pathological. For the classifier, most of the previous investigations have used support vector machines (SVMs) [5], [13], [16], [22]. In addition to SVMs, other algorithms such as artificial neural networks, decision trees, and variants of recurrent neural network (RNN) have also been used as classifiers in the study area [13], [28], [32]–[34]. A review of various techniques considered for both parts is given in [5]. Recently, a few studies have investigated pathological voice detection using end-to-end systems [15], [35]–[37]. In end-to-end systems, deep neural networks are trained to predict labels directly either from the raw time-domain voice signal or from the mel spectrogram [15]. However, unlike in many speech technology areas such as audio tagging [30] and speech synthesis [31], end-to-end systems have not been widely used in the field of voice pathology detection due to the lack of sufficient data. The data scarcity is an inherent problem in the study area of pathological voice because data is collected from patients whose condition might be so weak that only short recordings are possible. Hence, due to the lack of vast amounts of training data that is needed to train end-to-end systems, the traditional pipeline system is still a valid choice in voice pathology detection.

In the traditional pipeline system, cepstral features, particularly MFCCs, are most popular and they have been shown to perform comparably to or better than many other feature types [16], [44], [45]. The cepstral domain representation has the advantage that the features are less correlated, which is advantageous in the efficient implementation of ML classifiers. In addition, cepstral features can be computed without estimating the fundamental frequency (F_0) of the voice signal, which is beneficial compared to many perturbation features, complexity features, and glottal features where the extraction of F_0 is needed. Moreover, cepstral features do not need the extraction of glottal closure instants, which is needed, for example, in parameterizing voice pathologies with glottal measures. Among various existing cepstral features, MFCCs can be regarded as the *defacto* standard feature set in the area of voice pathology detection. MFCCs have also been widely used as default reference features in many areas outside pathology detection (like speaker recognition [38], speech spoof detection [39], speech mode classification [40], etc.). Moreover, MFCC features are widely included in larger generic feature sets (such as the openSMILE feature set [41], GeMAPS feature set [42], and ComParE feature set [43]) to capture vocal tract information from speech and voice signals. Previous studies in voice pathology detection have computed MFCCs almost exclusively from the default acoustic input, the voice signal (i.e., the pressure signal recorded by the microphone), to effectively capture vocal tract information [5], [13]. In a few recent studies, however, an alternative way of computing MFCCs has been investigated by first separating the voice signal into the glottal source and vocal tract using glottal inverse filtering (GIF) and then computing the MFCC feature set from the time-domain glottal source signal [16], [17]. Since the latter way of computing MFCCs focuses on the glottal source, this way to compute MFCCs is justified for use as a feature extraction method, particularly in the detection of voice pathologies, such as vocal nodules and dysphonia, which affect the vocal folds. In [16], the two ways of computing the MFCC feature set were compared in the detection of voice pathologies. The results showed that the best detection performance was achieved by combining the MFCCs computed from the voice signal with the MFCCs computed from the glottal source signal. It should be emphasized, however, that the second component estimated by GIF, the vocal tract transfer function, has not been used in the MFCC computation in any of the previous experiments reported in [16], [17].

MFCCs are computed by filtering the input signal with the perceptually motivated mel filterbank [48]. When the mel filterbank is replaced with a linear filterbank, the cepstral coefficients are termed linear-frequency cepstral coefficients (LFCCs). Despite the popularity of the MFCC features in most voice and speech classification studies, LFCCs have been shown to perform better than MFCCs in areas such as speaker recognition and the detection of spoofing [38], [39]. However, to the best of our knowledge, there are no previous studies reporting how LFCCs perform in voice pathology detection compared to MFCCs.

In the current study, a systematic comparison between different cepstral coefficient (CC) feature sets in the detection of pathological voices is investigated. Given the recent results reported in [16] indicating that combining the MFCCs computed from the voice signal with the MFCCs computed from the glottal source improves performance of voice pathology detection, the *first* aim of the current study is to understand how CCs extracted separately from *both* the glottal source *and* the vocal tract perform in the detection task compared to the CCs, which are extracted in a conventional manner from the voice signal. We hypothesize that by first separating the acoustic voice signal into the glottal source and vocal tract and by extracting CCs separately from both components, voice pathologies can be detected with better accuracy compared to extracting CCs from the voice microphone signal where contributions of the source and tract are merged [50]. In order to study this topic, voice signals need to be separated prior to the computation of CCs into glottal source signals and digital filters representing the vocal tract. In the current study, this source-filter separation is conducted using a GIF algorithm called the quasi-closed phase (QCP) method [47]. The QCP method was selected as the GIF method since its performance was shown in [47] to be better than that of four state-of-the-art techniques, for both modal and non-modal phonation types. The *second* aim of the study was to investigate the effect of using the mel-spaced vs. the linear-spaced filterbank (referred to in brief as the mel and linear filterbanks, respectively) in the CC computation. This aim is justified because despite the fact that the use of the mel filterbank is well motivated in areas such as automatic speech recognition [48] where modeling of human sound perception is essential, the involvement of a perceptually motivated mel filterbank in the CC computation might be questioned for signals such as the glottal source where phonemic vocal tract cues are absent. The experimental evaluations were conducted using voice signals representing the vowel /a/ from six databases (described in Section III-A). The detection performances were separately evaluated with SVM, which is the most popular classifier in voice pathology detection. As per our knowledge, this is the first voice pathology detection study investigating in detail the effect of extracting CCs separately from both the voice source and the vocal tract (rather than from the voice signal alone) and comparing the effect of the filterbank (mel vs. linear) in the CC computation.

The paper is organized as follows. Section II describes the extraction of cepstral features from the acoustic voice signal and from the glottal source and vocal tract computed by GIF. Section III describes the pathology databases studied and the classifier, which was used in the detection experiments. The results are reported in Section IV by also including comparisons with existing techniques. Finally, the conclusions of the study are drawn in Section V.

II. FEATURE EXTRACTION

In order to achieve the two aims of the study described in the previous section, six different cepstral feature sets were first computed as shown in Figure 1. The computational details of these CC feature sets are described in this section by dividing the presentation into two parts according to the first aim of the study: Section II-A describes the CC sets that were computed using the conventional approach, where the input to the cepstral computation is the voice signal, and Section II-B describes the CC sets that were extracted from the glottal source signal and from the vocal tract estimated by the QCP method. In order to achieve the second aim of the study, cepstral feature sets described in Section II-A and Section II-B were built using both the mel and the linear filterbank.

A. CCs COMPUTED FROM THE VOICE SIGNAL

In order to parameterize the voice signal, 13-dimensional cepstral vectors (including the 0^{th} coefficient) were computed using 30 ms Hamming-windowed frames with a 5 ms shift. From static coefficients, delta (Δ) and double-delta ($\Delta\Delta$) coefficients were also derived resulting in a 39-dimensional cepstral feature vector. Both the mel filterbank and the linear filterbank were used and the corresponding CC feature sets are referred to as the voice_MFCC and voice_LFCC, respectively.

B. CCs COMPUTED FROM THE OUTPUTS OF GIF

GIF refers to the technique of estimating the glottal volume velocity waveform from the voice/speech (pressure) signal recorded by microphone. Several GIF methods have been proposed in literature (for a review of GIF, see [46]). As mentioned earlier, this study utilizes the QCP algorithm [47] as the GIF method. In the QCP method, a parametric vocal tract model is first computed using an all-pole modeling method called weighted linear prediction (WLP) [47], [49] (More details will be described in Section II-B.1). Then, the input acoustic speech signal is inverse filtered with the computed vocal tract transfer function to obtain the estimate of the glottal flow waveform. Figure 2 shows the two outputs of QCP for a healthy speaker and a speaker with dysphonia. In this work, CCs are computed from both the estimated glottal flow and from the estimated vocal tract filter as described next in Sections II-A.1 and II-A.2 respectively.

1) CCs COMPUTED FROM THE GLOTTAL SOURCE

The estimated time-domain glottal source signals were parameterized using the similar cepstral computation as described for the voice signal in Section II-A, except that the input was the glottal source. The CC feature sets obtained by using the mel and linear filterbank are referred to as the source_MFCC and source_LFCC, respectively.

2) CCs COMPUTED FROM THE VOCAL TRACT

When cepstral features are computed from the acoustic voice signal or from the glottal source waveform using the MFCC/LFCC pipeline, the fast Fourier transform (FFT) is used as the spectral estimation method. While the CCs computed from the glottal source (i.e., the source_MFCC and source_LFCC described in Section II-B.1) represent information which originates from the voice excitation generated by the vocal folds, the CCs computed from the voice signal (i.e., the voice_MFCC and voice_LFCC described in Section II-A) carry information that is brought about by both the glottal excitation and the vocal tract [50]. This means that the FFT-based CCs computed from the voice signal might not be able to capture the vocal tract characteristics effectively in the case of voice pathologies that affect the vocal tract system. Therefore, it is justified to extract cepstral feature sets that focus on the vocal tract alone by computing the cepstrum using the parametric WLP all-pole spectrum of the vocal tract model estimated by QCP. WLP is a modified LP method that has been proposed to downgrade the prominent effect of the glottal source in the computation of the vocal tract transfer function [47], [49]. The effect of the glottal source is downgraded in WLP by using a specifically selected temporal weighting function, called the attenuated main excitation (AME) function [49]. The AME function decreases the effect of voice samples in the vicinity of glottal closure in the computation of the autocorrelation function in all-pole modeling. Hence, the use of WLP removes the source information from the cepstrum, which leads to a better

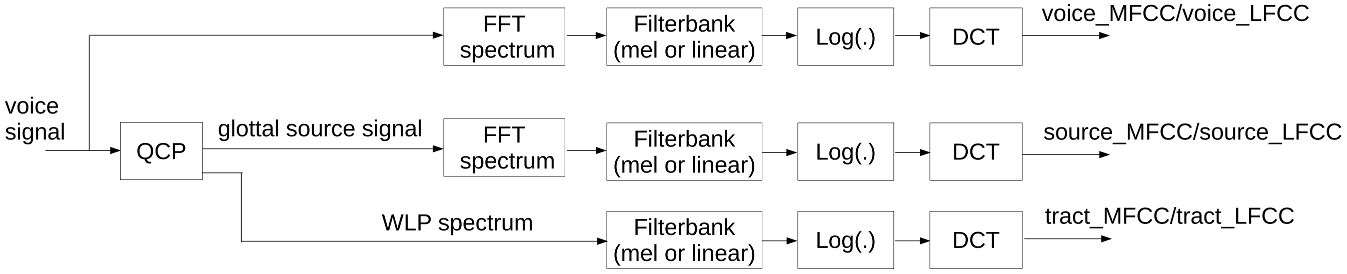

FIGURE 1. A flow diagram depicting the extraction of different cepstral feature sets studied in this article. Top path: extracting cepstral features from the voice signal. Middle path: extracting cepstral features from the glottal source signal estimated by QCP analysis. Bottom path: extracting cepstral features from the WLP vocal tract spectrum computed by QCP analysis. DCT represents discrete cosine transform. Two filterbanks (mel vs. linear) are used in each path.

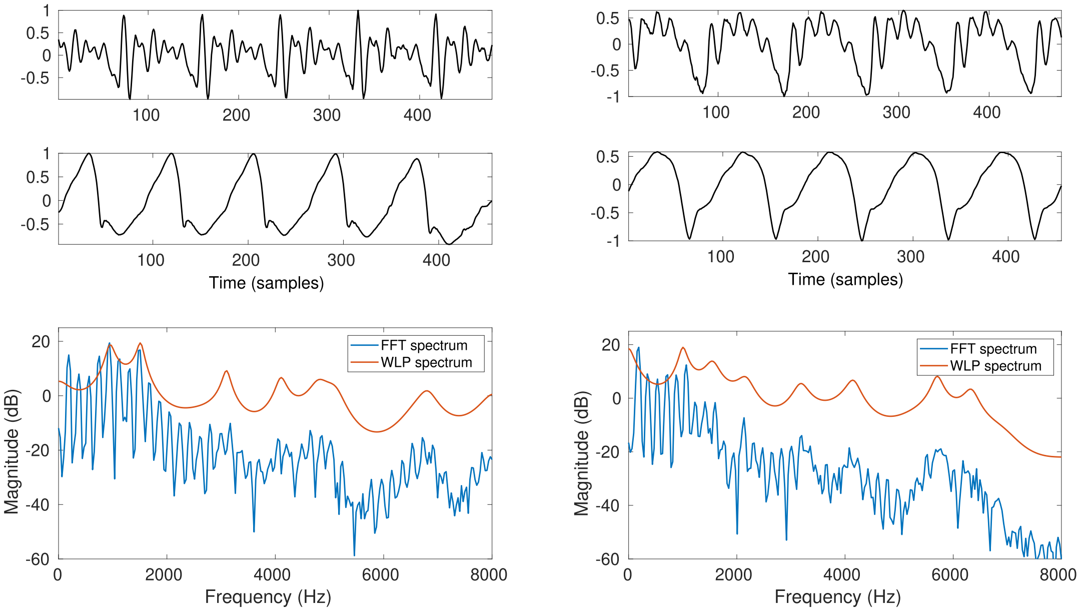

FIGURE 2. An illustration of GIF outputs for a healthy speaker (left panels) and a speaker with dysphonia (right panels). The top panels show the time-domain voice signals of the vowel /a/, the middle panels show the corresponding glottal flow waveforms, and the bottom panels show the corresponding WLP spectra of the vocal tract. The FFT spectra of the voice signals are shown as reference in the bottom panels.

representation of vocal tract information. The bottom panels of Figure 2 compare the FFT spectra computed from the voice signal and the WLP spectra of the vocal tract computed using the QCP method. The spectrum estimated with FFT displays a harmonic structure, which is conventionally smoothed by the perceptually motivated mel filterbank. However, the perceptual smoothing is only partially efficient at removing the harmonic structure present in voice signals [52]. On the other hand, the WLP method computes a parametric spectral envelope model of the vocal tract without modeling the harmonic structure. The WLP model order was chosen to be $p = 24$, which (according to [47]) is a valid vocal tract filter order when data is sampled with 16 kHz as in the current study. 39-dimensional CCs were computed in a similar way as in Section II-A except that the FFT magnitude spectrum of the voice signal was replaced with the WLP all-pole spectrum of the vocal tract. The resulting CC features sets are referred to as tract_MFCC and tract_LFCC when the mel and linear filterbank was used, respectively.

III. EXPERIMENTAL PROTOCOL

A. CONSIDERED VOICE PATHOLOGY DATABASES

The accuracy of the different cepstral feature sets constructed was evaluated in pathological voice detection using voice signals representing the vowel /a/ from six voice databases: the Hospital Universitario Príncipe de Asturias (HUPA) [51] database, the Neurovoz [53] database, the PC-GITA [54] database, the hyperkinetic dysphonia (HkD) and laryngitis subsets of the Saarbrücken Voice Disorders (SVD) [55], [56] database, and the heart failure (HF) database [57]. Most previous studies in voice pathology detection are based on voice recordings of sustained vowels, particularly representing the vowel /a/ [5], [58], [59]. All the pathology detection experiments conducted in the current investigation also focus on voice signals representing the vowel /a/ from the above databases. The popularity of using recordings of signals representing the vowel /a/ in the study area is explained by the fact that this vowel sound is easy to pronounce for patients of all languages. Moreover, the acoustical analysis of voice signals representing the vowel /a/ is easy because the vowel's formants are clearly distinguishable and peaks are prominent [58]. In addition to the general justifications above, studying recordings of the vowel /a/ in the current study was motivated by the use of GIF in this study. It has been shown namely that the high value of the first formant (F1) in the vowel /a/ makes the estimation of the glottal flow waveform with GIF more accurate [60].

1) *The HUPA database:* This database contains sustained phonations of the vowel /a/ by 439 adult Spanish speakers (239 healthy and 200 pathological). Pathological voices contain a wide variety of organic pathologies such as nodules, polyps, oedemas, and carcinomas. The data was recorded using the Kay Computerized Speech Lab Analysis station 4300B with a sampling frequency of 50 kHz and with a resolution of 16 bits. A detailed description of the database can be found in [51].
2) *The Neurovoz database:* This corpus consists of voice signals representing the vowel /a/ produced by 110 Parkinsonian patients and by 93 healthy speakers whose mother tongue is Castillian Spanish [53]. The signals were recorded with a sampling frequency of 44.1 kHz. More details about the Neurovoz database can be found in [53].
3) *The PC-GITA database:* This corpus contains voice signals collected using a variety of speaking tasks by 50 Parkinsonian patients and 50 control speakers whose native language is Colombian Spanish [54]. The data was recorded using a dynamic omni-directional microphone (Shure, SM 63L) and sampled at 44.1 kHz with a resolution of 16 bits. For the purpose of this study, we considered voice signals representing the vowel /a/ from the database. There are three repetitions of the vowel for each speaker. Hence, a total of 150 healthy and 150 PD voice signals were taken from this dataset to be used in the current experiments.
4) *The SVD-HkD database:* The SVD database [55], [56] is a large repository of pathological speech comprising recordings of sustained phonations of the vowels /a/, /i/, and /u/ in normal, high, and low pitches, as well as with rising-falling pitch. In addition, the data contains recordings of the sentence "Guten Morgen, wie geht es Ihnen?" ("Good morning, how are you?"). The database was recorded from 2225 German speakers, of which 869 are healthy and 1356 pathological. The entire database contains a total of as many as 71 different pathologies including both functional and organic pathologies. The data was recorded with a sampling frequency of 50 kHz. For the current study, we used a part of the SVD database by selecting voices representing the vowel /a/ produced with normal pitch by patients suffering from hypokinetic dysphonia (HkD). The selected subset of the entire SVD database is referred to as the SVD-HkD database in the current study. In the SVD-HkD database, there are 213 healthy utterances and 213 utterances produced by dysphonic patients, respectively.
5) *The SVD-Laryngitis database:* We also created another subset of the SVD database for the purposes of the current study by choosing laryngitis as the second voice pathology from the voices of the SVD database. This data will be referred to as the SVD-Laryngitis database in the current study. The SVD-Laryngitis database consists of 140 healthy and 140 disordered voice signals of the vowel /a/ produced at normal pitch by all the speakers.
6) *The HF database:* This is a new speech database which was recorded in Finnish by the authors as part of a previous study [57]. This database includes speech recordings from 20 speakers with HF and 25 healthy controls. The patients were hospitalized for HF of any etiology, regardless of the left ventricular ejection fraction. Each speaker read the same Finnish text three times (the text-reading task) and produced one spontaneous speech. The speech data, sampled at 44.1 kHz, was recorded in doctor's practice rooms using a headset condenser microphone (DPA 4065-BL) and an AD converter (RME Babyface Pro). A linear phase FIR filter (cut-off frequency: 60 Hz) was used to remove the low-frequency noise picked up by the recording microphone. For the purpose of this study, for each speaker, we considered four segments of the vowel /a/ extracted from the middle recitation of the text-reading task. Hence, in total, the HF database consists of 100 healthy utterances and 80 HF utterances.

The databases were recorded using different sampling rates. In order to maintain uniformity, we re-sampled all the voice signals to correspond to the same sampling rate of 16 kHz.

B. SVM CLASSIFIER

Support vector machine (SVM) is the most popular classifier for voice pathology detection. In this study, we used SVM with a radial basis function (RBF) kernel. Experiments were conducted separately with each database by considering data from 2/3 of the speakers for training and remaining for testing. This was repeated for 20 iterations, each time building different train and test data. In each iteration, there was no overlap between speakers used in training and testing. The CC feature vectors extracted from all the frames of an utterance were averaged, yielding 39-dimensional utterance-level feature vectors. The training data were z-score normalized and the testing data were normalized by subtracting the mean and dividing by the standard deviation of the training sets for each feature. The detection accuracy was used as the performance measure. The accuracy was computed as the ratio of the number of correctly classified voice signals to the total number of voice signals. From the detection accuracies obtained at all iterations, the mean and standard deviation were computed as the final performance measures to be used in comparing the different feature sets.

TABLE 1. Comparison of detection accuracies obtained with the defacto MFCC feature set (voice_MFCC) and its variants.

	Accuracy (in (%), mean ± std)						
Feature set	HUPA	Neurovoz	PC-GITA	SVD-HkD	SVD-Laryngitis	HF	Overall
voice_MFCC	72.89 ± 2.50	74.39 ± 2.94	**78.64 ± 3.53**	72.41 ± 1.35	69.48 ± 2.35	88.76 ± 2.82	**76.09 ± 2.58**
source_MFCC	**75.85 ± 1.52**	62.10 ± 3.02	68.67 ± 3.29	70.57 ± 2.23	**71.80 ± 2.08**	85.21 ± 2.50	72.36 ± 2.44
tract_MFCC	67.97 ± 3.13	**75.98 ± 3.62**	75.38 ± 2.39	72.91 ± 2.09	68.76 ± 2.27	**91.43 ± 3.12**	75.40 ± 2.77
voice_LFCC	71.69 ± 2.83	74.73 ± 1.88	75.72 ± 2.75	72.35 ± 1.88	69.41 ± 2.37	84.40 ± 2.63	74.71 ± 2.39
source_LFCC	70.80 ± 1.80	60.77 ± 3.15	71.11 ± 2.85	69.65 ± 2.93	70.29 ± 3.02	83.66 ± 2.49	71.04 ± 2.70
tract_LFCC	68.44 ± 2.80	73.44 ± 2.38	73.17 ± 3.23	**74.76 ± 1.54**	68.63 ± 2.35	84.82 ± 4.60	73.87 ± 2.81

TABLE 2. Detection accuracies obtained with combined feature sets.

	Accuracy (in (%), mean ± std)						
	Without feature selection						
Feature set	HUPA	Neurovoz	PC-GITA	SVD-HkD	SVD-Laryngitis	HF	Overall
source_MFCC + tract_MFCC	76.25 ± 1.77	71.97 ± 4.08	81.47 ± 2.63	**75.77 ± 1.37**	68.12 ± 3.42	93.85 ± 2.01	77.90 ± 2.55
source_MFCC + tract_LFCC	77.35 ± 3.06	75.41 ± 2.95	79.17 ± 2.23	73.81 ± 1.95	67.73 ± 3.45	93.08 ± 1.67	77.75 ± 2.55
source_LFCC + tract_LFCC	73.13 ± 2.12	73.62 ± 3.13	81.09 ± 1.70	72.16 ± 2.35	70.04 ± 2.67	94.13 ± 2.45	77.36 ± 2.40
source_LFCC + tract_MFCC	74.25 ± 1.90	72.46 ± 3.45	81.28 ± 2.03	72.82 ± 2.87	71.52 ± 2.07	93.52 ± 1.84	77.64 ± 2.36
	With feature selection						
source_MFCC + tract_MFCC	77.11 ± 1.46	69.95 ± 2.59	**82.03 ± 2.67**	75.66 ± 2.48	70.49 ± 2.17	**95.44 ± 2.74**	78.44 ± 2.35
source_MFCC + tract_LFCC	**78.86 ± 2.15**	**77.04 ± 3.61**	77.57 ± 3.94	74.73 ± 1.34	**74.79 ± 2.90**	91.57 ± 2.23	**79.09 ± 2.70**
source_LFCC + tract_LFCC	75.52 ± 2.23	72.89 ± 3.37	78.13 ± 2.94	72.01 ± 2.18	69.22 ± 1.79	89.29 ± 2.54	76.17 ± 2.51
source_LFCC + tract_MFCC	72.37 ± 1.62	74.60 ± 2.34	81.97 ± 1.95	72.62 ± 2.49	72.83 ± 2.28	90.91 ± 3.41	77.55 ± 2.35

IV. RESULTS

The experiments were carried out using the SVM classifier with the *individual* cepstral feature sets described in Section II as well as *combined* feature sets which were obtained by merging the CCs computed from the glottal source (described in Section II-B.1) and from the vocal tract (described in Section II-B.2). The latter analysis was included to study the potential existence of complementary information between the source-based and the vocal tract-based cepstral feature sets. Since the individual source-based and tract-based cepstral feature vectors were all of the same dimension of 39, their combinations yielded in principle cepstral feature vectors whose dimension was doubled, that is, 78. Therefore, the feature selection was used for the combined feature sets (as will be described in Section IV-B) and the accuracy is reported both with and without feature selection for the combined sets. Tables 1 and 2 show the results separately for the individual and combined feature sets. The main results reported in these tables are discussed in this section by separating the treatment in two sections: the results obtained using the individual CC feature sets are discussed in Section IV-A and the results obtained using the combined source/tract CC feature sets are discussed in Section IV-B.

A. DETECTION ACCURACY OBTAINED WITH THE INDIVIDUAL FEATURE SETS

By first comparing the mel filterbank-based individual feature sets, the following observations can be made from Table 1: voice_MFCC gave better accuracy compared to both source_MFCC and tract_MFCC only in one database (PC-GITA) whereas in the remaining databases, voice_MFCC was outperformed by either source_MFCC (HUPA and SVD-Laryngitis) or tract_MFCC (Neurovoz, SVD-HkD and HF). The source_MFCC gave the highest accuracy among all the six individual feature sets for HUPA and SVD-Laryngitis. In case of the Neurovoz and HF databases, the tract_MFCC gave better accuracies compared to all other individual feature sets.

By similarly comparing the linear filterbank-based feature sets, it can be seen from Table 1 that the accuracy of voice_LFCC was comparable or better than that of both source_LFCC and tract_LFCC for all databases except for the SVD-HkD and SVD-Laryngitis databases. The tract_LFCC and source_LFCC gave the highest accuracy among the six individual feature sets for SVD-HkD and SVD-Laryngitis, respectively. By comparing between the two filterbanks (mel vs. linear) used in the computation of voice-based, source-based, and tract-based feature sets, the following observations can be made. The accuracy achieved with voice_LFCC was better than or comparable to that of voice_MFCC in three databases (Neurovoz, SVD-HkD, SVD-Laryngitis). However, the accuracy obtained with source_LFCC was inferior to that given by source_MFCC in all databases except for the PC-GITA database. The tract_LFCC set gave comparable or better accuracy than tract_MFCC in three databases (HUPA, SVD-HkD, SVD-Laryngitis). Overall, for the considered databases, the results indicate that the use of the linear-frequency filterbank has a comparative advantage over the mel-frequency filterbank in extracting CCs from the voice signal and from the vocal tract for detecting voice pathologies. In the case of the glottal source, however, the CCs extracted using the mel filterbank gave better accuracies than those extracted with the linear filterbank. It is worth emphasizing that the voice_MFCC set, which represents the *defacto* feature set in the study area, gave the best accuracy among the six individual CC feature sets only in one database (PC-GITA). However, voice_MFCC provided the

TABLE 3. Detection accuracies for four existing feature extraction techniques (MFCC, LPCC, CQCC and PLPCC) and for the combination source_MFCC + tract_LFCC (denoted by Comb-CC). The four existing feature sets were all computed using the voice signal as input to the cepstral computation. Two classifiers (SVM and 1-D CNN) were used and the test data was either clean or corrupted by traffic noise using three SNR categories.

	SVM				1-D CNN			
	10 dB	20 dB	30 dB	Clean	10 dB	20 dB	30 dB	Clean
MFCC	72.22	**76.09**	80.31	82.37	**74.22**	**77.39**	83.31	88.37
LPCC	68.44	71.52	74.22	79.91	70.44	73.52	79.22	84.91
CQCC	65.76	69.58	71.14	74.66	67.76	70.58	74.14	78.66
PLPCC	**73.11**	75.84	79.78	83.65	73.11	76.44	**84.78**	89.15
Comb-CC	69.33	74.76	**82.62**	**85.83**	70.33	75.76	84.52	**90.43**

best performance (76.09%) in terms of mean accuracy averaged over all six databases (as seen from the last column in Table 1). In the next section, the CC sets computed separately from the source and vocal tract are *combined* and the combined sets are compared to the voice-based CCs which, in principle, carry mainly vocal tract information but also some source information.

B. DETECTION ACCURACY OBTAINED WITH THE COMBINED FEATURE SETS

The glottal source-based and the vocal tract-based CCs were combined to analyze complementary information among these feature sets, and to compare their performance with the voice signal-based CCs reported in Section IV-A. It should be noted that both the combined CCs (i.e., the combination of source_MFCC and tract_MFCC and the combination of source_LFCC and tract_LFCC) and the individual voice-based CCs (i.e., voice_MFCC and voice_LFCC) carry vocal tract as well as glottal source information. The dimension of the individual voice signal-based CC sets is 39 and that of the combined CC set is 78. Therefore, to perform a fair comparison between the CC feature sets of different dimensions, we reduced the size of the combined CC sets to 39 by using the non-parametric neighborhood component analysis (NCA) feature selection technique [61]. The NCA was implemented in the current study using MATLAB. The results obtained for the six databases with combined CCs are shown in Table 2. From the first column (HUPA), it can be seen that the combination source_MFCC and tract_MFCC/tract_LFCC and their reduced versions provided better accuracies compared to any of the six individual sets discussed in Section IV-A. The results shown in the fourth column (SVD-HkD) indicate that the combined sets (source_MFCC + tract_MFCC and source_MFCC and tract_LFCC) and their reduced counterparts perform comparably to or better than the best individual feature sets. This indicates that the reduced sets of the combinations of MFCC extracted from glottal source with MFCC/LFCC extracted from vocal tract (estimated using GIF) are more effective than the CCs extracted from the voice signal in the detection of dysphonia. The second and the third columns show the results obtained for the two PD datasets (Neurovoz and PC-GITA). In this case, the reduced versions of the combinations source_MFCC + tract_LFCC and source_LFCC + tract_MFCC perform comparably or better than voice_MFCC or voice_LFCC. The results obtained for the SVD-Laryngitis dataset (the fifth column) indicate that the highest accuracy was achieved by combining source_LFCC with tract_MFCC. However, among the reduced feature sets, the combination of source_MFCC and tract_LFCC gave the best detection accuracy. The reduced sets of all the combinations performed comparably or better than the voice signal-based CCs. In the case of the HF database (the sixth column), the results indicate that the full combined sets of the source-based and tract-based CCs as well as their reduced versions outperformed the voice-based CCs. The overall mean accuracy (average accuracy across all the databases) of the combined sets and their reduced versions is better than that of the individual voice signal-based CCs. Finally, the following important observation related to the two aims of the study can be made by comparing accuracies obtained with the *defacto* CC feature set (i.e., voice_MFCC shown by the first row in Table 1) to those obtained with the reduced CC set, which combines source_MFCC and tract_LFCC (i.e., the accuracies shown by the sixth row in Table 2): the latter feature set was shown to give better detection accuracy in all the databases except for PC-GITA. Furthermore, this feature set also gave the highest overall mean accuracy (79.09%) compared to other individual and combined feature sets.

C. COMPARISON WITH EXISTING TECHNIQUES

In the previous sub-sections, we compared the source-based and vocal tract-based cepstral features and their combinations using the voice-signal based cepstral features as the reference. The results in Table 2 show that the combination source_MFCC + tract_LFCC yielded the best overall detection accuracy. In this sub-section, this best feature combination will be compared to other existing feature extraction techniques. The combination source_MFCC + tract_LFCC will be shortly referred to as Comb-CC. Since different feature extraction techniques have been developed in many studies over the past decades, a large number of potential reference techniques is available. These include, for example, perceptual linear prediction features (used, e.g., in [62]), discrete wavelet transform features (used, e.g., in [63]), empirical mode decomposition features (used, e.g., in [64]), and non-negative matrix factorization-based features (used, e.g., in [65]). For more details about various existing feature extraction methods, the interested reader is referred to the

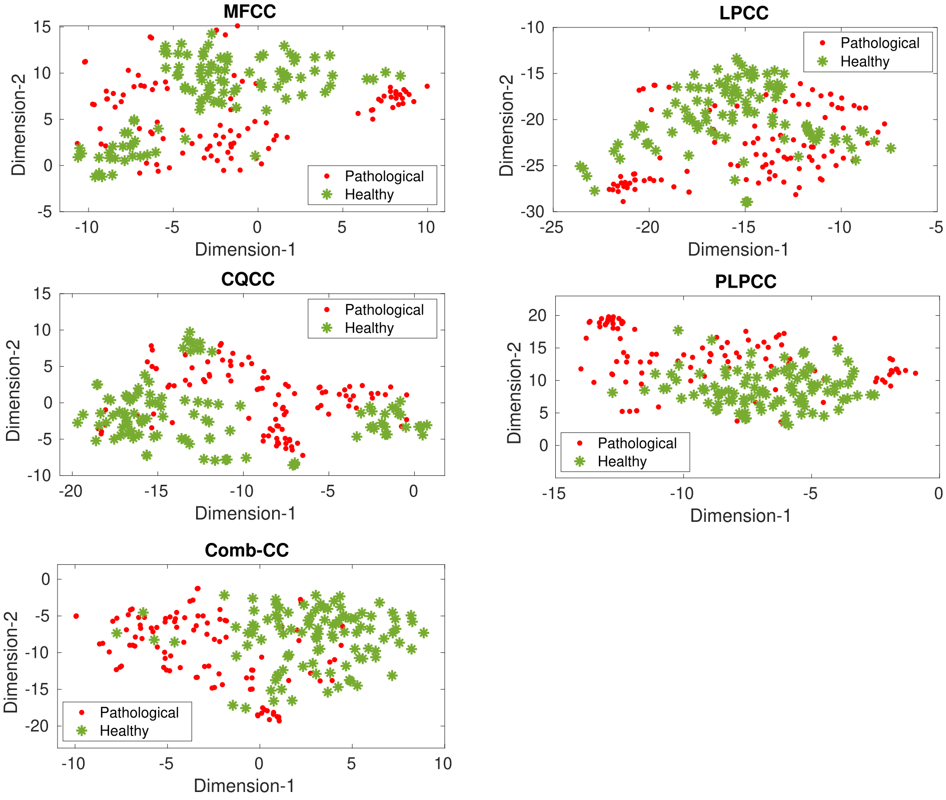

FIGURE 3. 2-D t-SNE visualization results plotted with 100 healthy and 100 pathological voice samples randomly selected from the clean test data set.

following tables in three reviews in voice pathology detection: Table 1 in [67], Table 2 in [5], and Table 1 in [66]. For comparison, we selected four existing features extraction techniques which have been widely used in the study area of pathology detection and which are all based on the computation of cepstral features. The selected four cepstral feature techniques are MFCCs, linear prediction cepstral coefficients (LPCCs), constant Q cepstral coefficients (CQCCs), and perceptual linear prediction cepstral coefficients (PLPCCs). It is worth emphasising that all the selected four existing techniques use the voice signal as input to the cepstrum computation whereas Comb-CC uses both the glottal source and the vocal tract as inputs to the cepstrum computation. In addition, the detection experiments are computed in this section using two classifiers: SVM and one-dimensional convolutional neural network (1D-CNN). The same SVM used for experimentation in the previous section is considered. The 1D-CNN architecture consists of 5 layers in total: two convolutional layers with 16 filters of length 3, alternating with max-pooling layers of pooling size 2, and one fully connected layer comprising 256 units. The input length of this model is 39 samples, which corresponds to the dimension of the CC feature vector. We adopted rectified linear units as the activation function in all layers. Note that all the convolutions and pooling operations used are one-dimensional. The voice samples from the six databases are combined to form a single dataset by considering only one voice sample per speaker from all the databases. Data from 70% of speakers is used for training and the remaining data is used for testing. The data from the test speakers is unseen during the training phase. To increase the training data, we corrupted the clean training data with additive noise in three different signal-to-noise (SNR) conditions (10 dB, 20 dB and 30 dB) using realistic non-stationary traffic noise. The (clean) training data was then augmented with the corrupted data. This procedure gave a 4-fold increase in the amount of the training data and may help to incorporate noise robustness of the model. The test data was also corrupted in the similar manner.

The results of comparing Comb-CC with the four existing techniques are shown in Table 3. From the table, it can be seen that the CQCC set showed the lowest detection accuracy both with the SVM and 1D-CNN classifier. The accuracy given by Comb-CC is lower compared to PLPCC and MFCC at the lowest SNR level (SNR = 10 dB). This is due to the sensitivity of GIF analysis to noise [46]. At the higher SNR levels (20 dB and 30 dB), the performance of Comb-CC is comparable or better than that of PLPCC and MFCC. The Comb-CC features outperform all other considered features in the case of clean voice signals for both the SVM and 1D-CNN classifier.

Figure 3 shows the t-distributed stochastic neighbor embedding (t-SNE) visualization plots of various cepstral features for 100 pathological and 100 normal voices randomly selected from the clean test data. The t-SNE is a non-linear dimensionality reduction algorithm used for exploring high-dimensional datasets by mapping to two or more dimensions suitable for human observation [68]. In Figure 3, the t-SNE displays the 39-dim cepstral features using two dimensions. From the figure, it can be seen that the two classes are clustered much better with the proposed Comb-CC features compared to the other features.

V. CONCLUSION

In the automatic detection of voice pathology, traditional pipeline systems based on using a separate feature extraction stage and a separate classification stage is still a valid system architecture, despite the fact that modern end-to-end systems provide excellent detection accuracy. One of the strongest benefits of traditional pipeline systems, particularly those based on the SVM classifier, is their good performance in scenarios where only little training data is available. Furthermore, it is difficult for the user (e.g. the clinician) to gain knowledge about the underlying reasons why a certain detection decision was made by the end-end network. Therefore, many studies have been published in automatic detection of voice pathologies by using SVM-based traditional pipeline systems [13], [16], [17], [58], [59]. These studies have almost exclusively used MFCC features as the method to express the voice signal in a parametric form, either alone or as the *defacto* reference method. Moreover, a few recent works [16], [17] have shown that the detection of certain pathologies benefits from combining the traditional voice signal-based MFCCs with the MFCCs computed from the glottal source waveforms estimated by GIF. However, none of the previous works in the study area have investigated the possibility of also using the other output of GIF, the estimated vocal tract filter, in the computation of cepstral features. In addition, there are no studies in automatic detection of voice pathologies comparing the detection performances between CCs derived using the mel and linear filterbanks. Therefore, the aims of the current study were to compare the performance of different cepstral feature sets in the automatic detection of voice pathologies by varying both the input of the CC computation (the voice signal vs. the glottal source vs. the vocal tract) and the type of filterbank (mel vs. linear). A total of six voice pathology databases were used and this data represented four pathologies (dysphonia, Parkinson's disease, laryngitis and heart failure).

The experiments showed that the voice signal-based CCs provided better results than the other individual feature sets only for one database (PC-GITA). For the other databases, the CCs computed from either the glottal source or vocal tract performed better than the voice signal-based CCs. The results also indicate that the use of the linear filterbank has an advantage over the mel filterbank in the computation of CCs particularly from the vocal tract. Most importantly, the results indicate that the best detection accuracy averaged over all six databases was obtained by applying feature selection to the combination of the mel filterbank-based CCs computed from the glottal source and linear filterbank-based CCs computed from the vocal tract. This combination of CC features gave an improvement of 3% in overall accuracy when compared to conventional MFCC features. Furthermore, the results obtained with the combined dataset show that the combination of CCs gave comparable or better performance than the existing features for clean speech and for speech of high SNR. This result was obtained by both the SVM and 1D-CNN classifier. Hence, this study shows encouraging results which indicate that the accuracy of the traditional two-stage system based on the *defacto* cepstral feature extraction method, the computation of MFCCs from the voice signal, can be improved with a proper combination of the glottal source-based and vocal tract-based cepstral features. In the future, the study can be extended to other voice pathologies such as dysarthria and vocal nodules. Furthermore, in order to improve the detection accuracy given by the proposed source-based and tract-based cepstral features in noisy conditions, the use of noise-robust GIF methods (e.g. [69]) in the proposed cepstral computation approach will be studied.

REFERENCES

[1] L. Daudet, N. Yadav, M. Perez, C. Poellabauer, S. Schneider, and A. Huebner, "Portable mTBI assessment using temporal and frequency analysis of speech," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 2, pp. 496–506, Mar. 2017.

[2] J. R. Orozco-Arroyave, E. A. Belalcazar-Bolaños, J. D. Arias-Londoño, J. F. Vargas-Bonilla, S. Skodda, J. Rusz, K. Daqrouq, F. Hönig, and E. Nöth, "Characterization methods for the detection of multiple voice disorders: Neurological, functional, and laryngeal diseases," *IEEE J. Biomed. Health Informat.*, vol. 19, no. 6, pp. 1820–1828, Nov. 2015.

[3] N. R. Smith, L. A. Rivera, M. Dietrich, C.-R. Shyu, M. P. Page, and G. N. DeSouza, "Detection of simulated vocal dysfunctions using complex sEMG patterns," *IEEE J. Biomed. Health Informat.*, vol. 20, no. 3, pp. 787–801, May 2016.

[4] I. R. Titze and D. W. Martin, "Principles of voice production," *Acoust. Soc. Amer. J.*, vol. 104, p. 1148, Sep. 1998.

[5] F. T. Al-Dhief, N. M. A. Latiff, N. N. N. A. Malik, N. S. Salim, M. M. Baki, M. A. A. Albadr, and M. A. Mohammed, "A survey of voice pathology surveillance systems based on Internet of Things and machine learning algorithms," *IEEE Access*, vol. 8, pp. 64514–64533, 2020.

[6] H. Teager, "Some observations on oral air flow during phonation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 5, pp. 599–601, Oct. 1980.

[7] C. Rosenbek and L. L. LaPointe, "The dysarthrias: Description, diagnosis, and treatment," in *Clinical Management of Neurogenic Communicative Disorders*, D. F. Johns, Ed. Boston, MA, USA: Little, Brown Co, 1985, pp. 97–152.

[8] M. J. Aminoff, H. H. Dedo, and K. Izdebski, "Clinical aspects of spasmodic dysphonia," *J. Neurol., Neurosurg. Psychiatry*, vol. 41, no. 4, pp. 361–365, 1978.

[9] T. Zhang, Y. Shao, Y. Wu, Z. Pang, and G. Liu, "Multiple vowels repair based on pitch extraction and line spectrum pair feature for voice disorder," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 7, pp. 1940–1951, Jul. 2020.

[10] J. A. Stamford, P. N. Schmidt, and K. E. Friedl, "What engineering technology could do for quality of life in Parkinson's disease: A review of current needs and opportunities," *IEEE J. Biomed. Health Inform.*, vol. 19, no. 6, pp. 1862–1872, Nov. 2015.

[11] J. C. Vásquez-Correa, T. Arias-Vergara, J. R. Orozco-Arroyave, B. Eskofier, J. Klucken, and E. Nöth, "Multimodal assessment of Parkinson's disease: A deep learning approach," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 4, pp. 1618–1630, Jul. 2019.

[12] B. E. Sakar, M. E. Isenkul, C. O. Sakar, A. Sertbas, F. Gurgen, S. Delil, H. Apaydin, and O. Kursun, "Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings," *IEEE J. Biomed. Health Inform.*, vol. 17, no. 4, pp. 828–834, Jul. 2013.

[13] M. K. Reddy, P. Alku, and K. S. Rao, "Detection of specific language impairment in children using glottal source features," *IEEE Access*, vol. 8, pp. 15273–15279, 2020.

[14] P. Carding, "Voice pathology in the United Kingdom," *BMJ*, vol. 327, no. 7414, pp. 514–515, Sep. 2003.

[15] N. P. Narendra and P. Alku, "Glottal source information for pathological voice detection," *IEEE Access*, vol. 8, pp. 67745–67755, 2020.

[16] S. R. Kadiri and P. Alku, "Analysis and detection of pathological voice using glottal source features," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 2, pp. 367–379, Feb. 2020.

[17] M. Borsky, D. D. Mehta, J. H. Van Stan, and J. Gudnason, "Modal and nonmodal voice quality classification using acoustic and electroglottographic features," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 12, pp. 2281–2291, Dec. 2017.

[18] A. A. Dibazar, S. Narayanan, and T. W. Berger, "Feature analysis for automatic detection of pathological speech," in *Proc. 2nd Joint 24th Annu. Conf. Annu. Fall Meeting Biomed. Eng. Soc. Eng. Med. Biol.*, vol. 1, 2002, pp. 182–183.

[19] F. Rudzicz, "Phonological features in discriminative classification of dysarthric speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Taipei, Taiwan, Apr. 2009, pp. 4605–4608.

[20] P. Gómez-Vilda, R. Fernández-Baillo, V. Rodellar-Biarge, V. N. Lluis, A. Álvarez-Marquina, L. M. Mazaira-Fernández, R. Martínez-Olalla, and J. I. Godino-Llorente, "Glottal source biometrical signature for voice pathology detection," *Speech Commun.*, vol. 51, no. 9, pp. 759–781, Sep. 2009.

[21] V. Uloza, A. Verikas, M. Bacauskiene, A. Gelzinis, R. Pribuisiene, M. Kaseta, and V. Saferis, "Categorizing normal and pathological voices: Automated and perceptual categorization," *J. Voice*, vol. 25, no. 6, pp. 700–708, Nov. 2011.

[22] N. P. Narendra and P. Alku, "Dysarthric speech classification from coded telephone speech using glottal features," *Speech Commun.*, vol. 110, pp. 47–55, Jul. 2019.

[23] I. A. Rezek and S. J. Roberts, "Stochastic complexity measures for physiological signal analysis," *IEEE Trans. Biomed. Eng.*, vol. 45, no. 9, pp. 1186–1191, Sep. 1998.

[24] Y. Wu, C. Zhou, Z. Fan, D. Wu, X. Zhang, and Z. Tao, "Investigation and evaluation of glottal flow waveform for voice pathology detection," *IEEE Access*, vol. 9, pp. 30–44, 2021.

[25] C. Manfredi, "Adaptive noise energy estimation in pathological speech signals," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 11, pp. 1538–1543, Nov. 2000.

[26] D. G. Childers and K. SungBae, "Detection of laryngeal function using speech and electroglottographic data," *IEEE Trans. Biomed. Eng.*, vol. 39, no. 1, pp. 19–25, Jan. 1992.

[27] M. D. O. Rosa, J. C. Pereira, and M. Grellet, "Adaptive estimation of residue signal for voice pathology diagnosis," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 1, pp. 96–104, Jan. 2000.

[28] J. I. Godino-Llorente and P. Gómez-Vilda, "Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 2, pp. 380–384, Feb. 2004.

[29] J. D. Arias-Londoño, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, and G. Castellanos-Domínguez, "Automatic detection of pathological voices using complexity measures, noise parameters, and mel-cepstral coefficients," *IEEE J. Trans. Bio-Med. Eng.*, vol. 58, no. 2, pp. 370–379, Feb. 2011.

[30] K. Xu, B. Zhu, Q. Kong, H. Mi, B. Ding, D. Wang, and H. Wang, "General audio tagging with ensembling convolutional neural networks and statistical features," *J. Acoust. Soc. Amer.*, vol. 145, no. 6, pp. EL521–EL527, Jun. 2019.

[31] L. Juvela, B. Bollepalli, V. Tsiaras, and P. Alku, "GlotNet—A raw waveform model for the glottal excitation in statistical parametric speech synthesis," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 6, pp. 1019–1030, Jun. 2019.

[32] L. A. Forero M., M. Kohler, M. M. B. R. Vellasco, and E. Cataldo, "Analysis and classification of voice pathologies using glottal signal parameters," *J. Voice*, vol. 30, no. 5, pp. 549–556, Sep. 2016.

[33] D. Hemmerling, J. R. Orozco-Arroyave, A. Skalski, J. Gajda, and E. Nöth, "Automatic detection of Parkinson's disease based on modulated vowels," in *Proc. INTERSPEECH*, San Francisco, CA, USA, 2016, pp. 1190–1194.

[34] A. Mayle, Z. Mou, R. Bunescu, S. Mirshekarian, L. Xu, and C. Liu, "Diagnosing dysarthria with long short-term memory networks," in *Proc. INTERSPEECH*, Graz, Austria, Sep. 2019, pp. 4514–4518.

[35] H. Wu, J. Soraghan, A. Lowit, and G. Di-Caterina, "A deep learning method for pathological voice detection using convolutional deep belief networks," in *Proc. INTERSPEECH*, Hyderabad, India, Sep. 2018, pp. 446–450.

[36] A. Rueda and S. Krishnan, "Augmenting dysphonia voice using Fourier-based synchrosqueezing transform for a CNN classifier," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brighton, U.K., May 2019, pp. 6415–6419.

[37] M. Alhussein and G. Muhammad, "Voice pathology detection using deep learning on mobile healthcare framework," *IEEE Access*, vol. 6, pp. 41034–41041, 2018.

[38] X. Zhou, D. Garcia-Romero, R. Duraiswami, C. Espy-Wilson, and S. Shamma, "Linear versus mel frequency cepstral coefficients for speaker recognition," in *Proc. IEEE Workshop Automat. Speech Recognit. Understand.*, Dec. 2011, pp. 559–564.

[39] M. R. Kamble, H. Tak, and H. A. Patil, "Amplitude and frequency modulation-based features for detection of replay spoof speech," *Speech Commun.*, vol. 125, pp. 114–127, Dec. 2020.

[40] K. Tripathi, M. K. Reddy, and K. S. Rao, "Multilingual and multimode phone recognition system for Indian languages," *Speech Commun.*, vol. 119, pp. 12–23, May 2020.

[41] F. Eyben, M. W. Öllmer, and B. Schuller, "openSMILE—The Munich versatile and fast open-source audio feature extractor," in *Proc. ACM Multimedia*, Florence, Italy, 2010, pp. 1459–1462.

[42] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Trans. Affective Comput.*, vol. 7, no. 2, pp. 190–202, Apr./Jun. 2015.

[43] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The INTERSPEECH 2016 computational paralinguistics challenge: Deception, sincerity & native language," in *Proc. INTERSPEECH*, Sep. 2016, pp. 2001–2005.

[44] J. A. Gómez-García, L. Moro-Velázquez, and J. I. Godino-Llorente, "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*, vol. 48, pp. 128–143, Feb. 2019.

[45] J. Monge-Álvarez, C. Hoyos-Barceló, P. Lesso, and P. Casaseca-de-la-Higuera, "Robust detection of audio-cough events using local Hu moments," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 1, pp. 184–196, Jan. 2019.

[46] P. Alku, "Glottal inverse filtering analysis of human voice production—A review of estimation and parameterization methods of the glottal excitation and their applications," *Sadhana*, vol. 36, no. 5, pp. 623–650, Oct. 2011.

[47] M. Airaksinen, T. Raitio, B. Story, and P. Alku, "Quasi closed phase glottal inverse filtering analysis with weighted linear prediction," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 3, pp. 596–607, Mar. 2014.

[48] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.

[49] P. Alku, J. Pohjalainen, M. Vainio, A.-M. Laukkanen, and B. H. Story, "Formant frequency estimation of high-pitched vowels using weighted linear prediction," *J. Acoust. Soc. Amer.*, vol. 134, no. 2, pp. 1295–1313, Aug. 2013.

[50] J. I. Godino-Llorente, P. Gomez-Vilda, and M. Blanco-Velasco, "Dimensionality reduction of a pathological voice quality assessment system based on Gaussian mixture models and short-term cepstral parameters," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 10, pp. 1943–1953, Oct. 2006.

[51] J. I. Godino-Llorente, V. Osma-Ruiz, N. Sáenz-Lechón, I. Cobeta-Marco, R. González-Herranz, and C. Ramírez-Calvo, "Acoustic analysis of voice using WPCVox: A comparative study with multi dimensional voice program," *Eur. Arch. Oto-Rhino-Laryngol.*, vol. 265, no. 4, pp. 465–476, Apr. 2008.

[52] S. Keronen, J. Pohjalainen, P. Alku, and M. Kurimo, "Noise robust LVCSR feature extraction based on stabilized weighted linear prediction," in *Proc. 13th Int. Conf. Speech Comput.*, St. Petersburg, Russia, Jun. 2009, pp. 221–225.

[53] L. Moro-Velazquez, J. A. Gomez-Garcia, J. I. Godino-Llorente, F. Grandas-Perez, S. Shattuck-Hufnagel, V. Yagüe-Jimenez, and N. Dehak, "Phonetic relevance and phonemic grouping of speech in the automatic detection of Parkinson's disease," *Sci. Rep.*, vol. 9, no. 1, pp. 1–16, Dec. 2019.

[54] 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. Lang. Resour. Eval. Conf. (LREC)*, 2014, pp. 342–347.

[55] M. Pützer and W. J. Barry, "Instrumental dimensioning of normal and pathological phonation using acoustic measurements," *Clin. Linguistics Phonetics*, vol. 22, no. 6, pp. 407–420, Jan. 2008.

[56] Institute of Phonetics, University of Saarland. (2010). *Saarbrücken Voice Database*. [Online]. Available: http://www.stimmdatenbank.coli.uni-saarland.de/

[57] M. K. Reddy, P. Helkkula, Y. M. Keerthana, K. Kaitue, M. Minkkinen, H. Tolppanen, T. Nieminen, and P. Alku, "The automatic detection of heart failure using speech signals," *Comput. Speech Lang.*, vol. 69, Sep. 2021, Art. no. 101205.

[58] G. Muhammad, G. Altuwaijri, M. Alsulaiman, Z. Ali, T. A. Mesallam, M. Farahat, K. H. Malki, and A. Al-Nasheri, "Automatic voice pathology detection and classification using vocal tract area irregularity," *Biocybern. Biomed. Eng.*, vol. 36, no. 2, pp. 309–317, 2016.

[59] R. Amami and A. Smiti, "An incremental method combining density clustering and support vector machines for voice pathology detection," *Comput. Electr. Eng.*, vol. 57, pp. 257–265, Jan. 2017.

[60] P. Alku, B. Story, and M. Airas, "Estimation of the voice source from speech pressure signals: Evaluation of an inverse filtering technique using physical modelling of voice production," *Folia Phoniatrica Logopaedica*, vol. 58, no. 2, pp. 102–113, 2006.

[61] W. Yang, K. Wang, and W. Zuo, "Neighborhood component feature selection for high-dimensional data," *J. Comput.*, vol. 7, no. 1, pp. 161–168, Jan. 2012.

[62] 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, Oct. 2017, Art. no. e0185613.

[63] I. Hammami, L. Salhi, and S. Labidi, "Voice pathologies classification and detection using EMD-DWT analysis based on higher order statistic features," *IRBM*, vol. 41, no. 3, pp. 161–171, Jun. 2020.

[64] T. Zhang, Y. Zhang, H. Sun, and H. Shan, "Parkinson disease detection using energy direction features based on EMD from voice signal," *Biocybern. Biomed. Eng.*, vol. 41, no. 1, pp. 127–141, Jan. 2021.

[65] B. Karan, S. S. Sahu, J. R. Orozco-Arroyave, and K. Mahto, "Non-negative matrix factorization-based time-frequency feature extraction of voice signal for Parkinson's disease prediction," *Comput. Speech Lang.*, vol. 69, Sep. 2021, Art. no. 101216.

[66] L. Moro-Velazquez, J. A. Gomez-Garcia, J. D. Arias-Londoño, N. Dehak, and J. I. Godino-Llorente, "Advances in Parkinson's disease detection and assessment using voice and speech: A review of the articulatory and phonatory aspects," *Biomed. Signal Process. Control*, vol. 66, Apr. 2021, Art. no. 102418.

[67] S. Hegde, S. Shetty, S. Rai, and T. Dodderi, "A survey on machine learning approaches for automatic detection of voice disorders," *J. Voice*, vol. 33, no. 6, pp. 947.e11–947.e33, 2019.

[68] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[69] N. P. Narendra, M. Airaksinen, B. Story, and P. Alku, "Estimation of the glottal source from coded telephone speech using deep neural networks," *Speech Commun.*, vol. 106, pp. 95–104, Jan. 2019.

MITTAPALLE KIRAN REDDY received the M.E. degree in communication systems from the SSN College of Engineering, Chennai, India, in 2014, and the Ph.D. degree in speech processing from the Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur, India, in 2019. From October 2014 to March 2018, he worked as a Senior Scientific Officer in the research project sponsored by the Department of Information Technology, India, undertaken by IIT Kharagpur. He is currently a Postdoctoral Researcher with the Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland. His research interests include speech processing, signal processing, pattern recognition, and machine learning.

PAAVO ALKU (Fellow, IEEE) received the M.Sc., Lic.Tech., and Dr.Sc. (Tech.) degrees from Helsinki University of Technology, Espoo, Finland, in 1986, 1988, and 1992, respectively. He was an Assistant Professor with Asian Institute of Technology, Bangkok, Thailand, in 1993, and the University of Turku, Finland, from 1994 to 1999. He is currently a Professor of speech communication technology with Aalto University, Espoo. His research interests include analysis and parameterization of speech production, statistical parametric speech synthesis, spectral modeling of speech, speech enhancement, and cerebral processing of speech.