Kadiri, Sudarsana; Alku, Paavo

Analysis and Detection of Pathological Voice using Glottal Source Features

Document Version
Peer reviewed version

Analysis and Detection of Pathological Voice using Glottal Source Features

Sudarsana Reddy Kadiri and Paavo Alku, *Senior Member, IEEE*

Abstract—Automatic detection of voice pathology enables objective assessment and earlier intervention for the diagnosis. This study provides a systematic analysis of glottal source features and investigates their effectiveness in voice pathology detection. Glottal source features are extracted using glottal flows estimated with the quasi-closed phase (QCP) glottal inverse filtering method, using approximate glottal source signals computed with the zero frequency filtering (ZFF) method, and using acoustic voice signals directly. In addition, we propose to derive mel-frequency cepstral coefficients (MFCCs) from the glottal source waveforms computed by QCP and ZFF to effectively capture the variations in glottal source spectra of pathological voice. Experiments were carried out using two databases, the Hospital Universitario Príncipe de Asturias (HUPA) database and the Saarbrücken Voice Disorders (SVD) database. Analysis of features revealed that the glottal source contains information that discriminates normal and pathological voice. Pathology detection experiments were carried out using support vector machine (SVM). From the detection experiments it was observed that the performance achieved with the studied glottal source features is comparable or better than that of conventional MFCCs and perceptual linear prediction (PLP) features. The best detection performance was achieved when the glottal source features were combined with the conventional MFCCs and PLP features, which indicates the complementary nature of the features.

Index Terms—Speech analysis, Pathological voice, Pathology detection, Glottal source features, Glottal flow waveform, Glottal inverse filtering.

I. INTRODUCTION

Speech is produced by exciting a time-varying vocal tract system that consists of various articulators (such as the tongue, jaw, lips) by a time-varying excitation signal. The main purpose of speech in communication is to convey a linguistic message. Apart from linguistic content, speech also contains rich information about the language and dialect as well as about the speaker's gender, age, emotions and state of health. This work studies pathological voice and compares it to normal voice using both analysis and detection (i.e., normal vs. pathological). Voice pathologies arise due to infections, physiological and psychogenic causes and due to vocal misuse that is prevalent in professions such as singers, teachers, and customer service representatives [1], [2]. Automatic detection of voice pathology enables an objective assessment and an early intervention for the diagnosis. A typical voice pathology detection system consists of two main stages: the first stage is the representation of the input acoustic speech signal (i.e.,

feature extraction) and the second stage is the classifier (i.e., normal vs. pathological decision). The main focus of the current study is on the first stage.

Feature sets used for voice pathology detection can be broadly classified into the following three categories [3], [4]: (1) perturbation measures, (2) spectral and cepstral measures, and (3) complexity measures. Perturbation measures capture the presence of aperiodicity and aspiration noise in the acoustic speech signal that occur due to irregular movement of the vocal folds and incomplete glottal closure. The popular parameters in this category are jitter and shimmer [5]–[11]. Jitter measures the short-term perturbations of the fundamental frequency (F_0) and shimmer measures the short-term perturbations in amplitude [5], [12]. Several variations of jitter (such as relative jitter, relative jitter average perturbation and jitter five-point period perturbation quotient) and shimmer (such as absolute shimmer, relative shimmer and shimmer three-point amplitude perturbation quotient) have been used for voice pathology detection [12], [13]. The estimation of these features depends on F_0, but accurate estimation of F_0 is known to be difficult in pathological voice [14], [15]. Even though many previous studies have investigated jitter and shimmer features, it is worth observing that these features are not included in the feature set recommended by the American Speech-Language-Hearing Association due to their lack of clinical voice utility [16]. For more details of the recommended acoustic measures, see Table 2 in [16]. Other popular perturbation measures that quantify the presence of aspiration noise include the harmonics-to-noise ratio (HNR) [17], [18], normalized noise entropy (NNE) [19], and glottal-to-noise excitation (GNE) ratio [20]–[22]. HNR is defined as the ratio between the harmonic component energy and the noise component energy. NNE is the ratio between the energy of noise and the total energy of the signal. GNE measures the correlation between Hilbert envelopes in different frequency bands of the acoustic speech signal.

Measures derived from spectrum and cepstrum have been used extensively for voice pathology detection because these methods are typically easy to compute and do not need the estimation of F_0 [15], [23], [24]. The popular features in this category are mel-frequency cepstral coefficients (MFCCs) [4], [15], [25] that utilize the principles of human auditory processing in the mel-scale and the decorrelating property of cepstrum. In addition, linear predictive cepstral coefficients (LPCCs) [21], [26], [27] and perceptual linear prediction (PLP) [4], [28] coefficients have been used in voice pathology detection. LPCCs capture the vocal tract system characteristics. PLP features are based on the modelling of the human au-

S.R. Kadiri and P. Alku are with the Dept. of Signal Processing and Acoustics, Aalto University, Finland, E-mail: {sudarsana.kadiri;paavo.alku}@aalto.fi. This study was funded by the Academy of Finland (project no. 284671, 312490).

ditory system using the Bark scale, equal loudness-level curve and intensity-to-loudness conversion [29]. Another popular feature in this category is the cepstral peak prominence (CPP) [30], [31]. A larger CPP value indicates a more prominent periodic structure of the signal. A variant of CPP that has been proposed by smoothing the cepstrum and the corresponding parameter is referred to as the smoothed CPP [32]. Studies [33], [34] have additionally used the average spectral energies in low-frequency and high-frequency bands. Features derived from time-frequency decomposition techniques such as adaptive time-frequency transform [35], [36], wavelet transform [37]–[40], modulation spectrum [41]–[43] and empirical mode decomposition [44] have also been investigated for voice pathology detection.

Complexity measures have been proposed to capture properties such as aperiodicity, non-linearity and non-stationarity present in the signal through estimators based on non-linear dynamic analysis [5], [45]–[50]. It is known that nonlinear phenomena are common in natural physiological systems such as speech production. Non-linear dynamic analysis characterizes the dynamic changes in voice pathologies that occur due to irregular and improper movement of the vocal folds. The popular parameters in this category are computed using the fractal dimension or the correlation dimension [28], [45], [51]–[53]. The complex measures investigated in several studies consist of the following: the largest Lyapunov exponent, the recurrence period density entropy, Hurst exponent, detrended fluctuation analysis, approximate entropy, sample entropy, modified sample entropy, Gaussian kernel sample entropy, fuzzy entropy, hidden Markov model (HMM) entropy and Shannon HMM entropy [38], [39], [54], [55]. These features capture the dynamic variants/invariants, long-range correlations, regularity or predictability information present in the signal.

It should be noted that the estimation of perturbation features and complexity features depends on the precise estimation of F_0 and the selection of the appropriate window length [14], [56]. On the other hand, extraction of spectral or cepstral features does not depend on F_0. In [13] and [57], it was found that voice pathology detection performance with spectral features (such as MFCCs and PLPs) alone is comparable or better than that given by perturbation and complexity features in sustained vowels and continuous speech. More details of the studies on pathological voice and various features used for voice pathology detection can be found in recent review articles [4], [13]. Regarding classifiers, several known techniques, such as kNN, GMM, LDA, HMM, ANN, CNN and SVM, have been used for pathological voice [25], [27], [40], [58]–[66]. Among the different classifiers, SVM has been found to be the most suitable classifier for voice pathology detection [67]. More details of various classifiers and machine learning techniques used for voice pathology detection can be found in the recent review published in [67].

Since voice pathologies affect the speech production mechanism, both the glottal source and the vocal tract system need to be represented and parameterized effectively in the analysis and detection of voice pathology. Existing studies have captured the vocal tract system characteristics effectively

by deriving spectral or cepstral features such as MFCCs and PLPs. However, there is little previous research on the systematic investigation of glottal source features for the analysis and detection of voice pathologies. In the few studies [68]–[71], authors have mainly exploited features that capture the specific glottal source characteristics such as HNR, GNE and spectral energies in low-frequency and high-frequency bands of the glottal source.

The current study presents a systematic analysis of glottal source features in normal and pathological voice and investigates their effectiveness in voice pathology detection. The glottal source features are derived from the glottal flow waveforms estimated using the quasi-closed phase (QCP) glottal inverse filtering method [72] and from the approximate glottal source signals computed by the zero frequency filtering (ZFF) method [73]. The glottal flow signals estimated using QCP are parameterized in terms of time-domain and frequency-domain features [74], [75]. The features derived from the ZFF method consist of the strength of excitation (SoE), energy of excitation (EoE), loudness measure and ZFF signal energy [76]. In addition to parameterizing glottal source waveforms computed by QCP and ZFF, we also use features which are derived directly from acoustic speech signals and which capture the specific property of the glottal source. These features are the maximum dispersion quotient (MDQ) [77], peak slope (PS) [78], cepstral peak prominence (CPP) [30], and Rd shape parameter [79], [80]. Further, we propose to derive MFCCs from the glottal source waveforms to effectively capture glottal source variations in pathological voice. In total, this results in five sets of glottal source features as follows.

- Time-domain and frequency-domain features derived from the glottal source waveforms estimated with the QCP method
- Features derived from the approximate glottal source waveforms computed by the ZFF method
- Features which are derived directly from acoustic voice signals and which capture the specific property of the glottal source
- MFCCs derived from the glottal flow waveforms estimated with the QCP method
- MFCCs derived from the approximate glottal source waveforms given by the ZFF method

Voice pathology detection experiments were carried out using two databases, the Hospital Universitario Príncipe de Asturias (HUPA) database [81], [82] and the Saarbrücken Voice Disorders (SVD) database [83], [84] that are considered the most reliable and standard databases for voice pathology detection [4], [13], [85]. We did not utilize the popular MEEI database because it suffers from the problems such as having different recording conditions between healthy and pathological voices (see, e.g., [3], [21], [47]). The conventional MFCC and PLP features, which were shown to be effective for voice pathology detection in [13], are used as the baseline features. Additionally, the complementary nature of the glottal source features is demonstrated, when the glottal source features are combined with the conventional MFCC and PLP features.

The paper is organized as follows. Section II describes the

two signal processing methods, QCP and ZFF, for deriving glottal source waveforms. The extraction of the glottal source features is discussed in Section III. Section IV presents the systematic analysis of the glottal source features for normal and pathological voice. Section V describes the extraction of MFCCs from the glottal source waveforms. Experimental protocol is discussed in Section VI, which includes the pathology databases, parameters used for feature extraction, baseline features used for comparison, details of the classifier and evaluation metrics. Results and discussion of the detection experiments are presented in Section VII. Finally, Section VIII summarizes the paper.

II. SIGNAL PROCESSING METHODS USED FOR DERIVING GLOTTAL SOURCE WAVEFORMS

This section describes the two signal processing methods used in the present study, the QCP glottal inverse filtering method [72] and the ZFF method [73], for the estimation of glottal source waveforms. It should be noted that QCP is based on the source-filter model of speech production but ZFF does not use the source-filter model. Hence, these two methods are expected to capture distinct information.

A. The quasi-closed phase (QCP) method

The QCP method [72] is a recently proposed glottal inverse filtering method for the automatic estimation of the glottal source waveform from speech. The method is based on the principles of closed phase (CP) [86] analysis which estimates the vocal tract model from few speech samples located in the CP of the glottal cycle using linear prediction (LP) analysis. In contrast to the CP method, QCP takes advantage of all the speech samples of the analysis frame in computing the vocal tract model. This is carried out using weighted linear prediction (WLP) analysis with the attenuated main excitation (AME) [87] waveform as the weighting function. The AME function is designed using glottal closure instants (GCIs) and the function attenuates the contribution of the open phase samples in the computation of the acoustic speech signal's covariance or autocorrelation function. This operation results in better estimates of the vocal tract transfer function $V(z)$. Finally, the estimate of the glottal flow waveform is obtained by inverse filtering the input acoustic speech signal with the vocal tract transfer function $V(z)$. The QCP method was shown to be better than four existing inverse filtering methods in the estimation of the glottal flow from modal and non-modal types of phonation [72]. This justifies the usage of QCP as a glottal inverse filtering method in the present study. A block diagram describing the steps involved in the QCP method is shown in Fig. 1.

B. The zero frequency filtering (ZFF) method

The ZFF method was proposed in [73] based on the fact that the effect of an impulse-like excitation (that occurs at the instant of glottal closure) is present throughout the spectrum including the zero frequency, while the vocal tract characteristics are mostly reflected in resonances at much
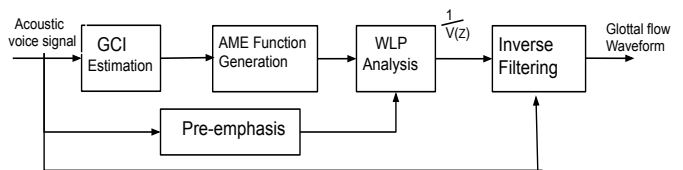

Fig. 1. Block diagram of the QCP method.

higher frequencies. In this method, the acoustic speech signal is passed through a cascade of two zero frequency resonators and the resulting signal is equivalent to integration of the signal four times. Hence, the output grows or decays as a polynomial function of time. The trend is removed by subtracting the local mean computed over the average pitch period at each sample and the resulting output signal is referred as the zero frequency filtered (ZFF) signal. In this study, we consider the ZFF signal as an approximate glottal source waveform.

The following steps are involved to derive the ZFF signal:

1) The acoustic voice signal ($s[n]$) is first differentiated as follows to remove any low-frequency trend

$$x[n] = s[n] - s[n-1]. \tag{1}$$

2) The differentiated signal is passed through a cascade of two zero frequency resonators (pair of poles on the unit circle along the positive real axis in the z-plane). This filtering can be expressed as follows

$$y_o[n] = \sum_{k=1}^{4} a_k y_o[n-k] + x[n], \tag{2}$$

where $a_1 = +4$, $a_2 = -6$, $a_3 = +4$, $a_4 = -1$. The resulting signal $y_o[n]$ is equivalent to integration (or cumulative sum in the discrete-time domain) of the acoustic voice signal four times, hence it approximately grows or decays as a polynomial function of time.

3) The trend in $y_o[n]$ is removed by subtracting the local mean computed over the average pitch period (derived using autocorrelation) at each sample. The resulting signal ($y[n]$) is called the ZFF signal and is computed as follows

$$y[n] = y_o[n] - \frac{1}{2N+1} \sum_{i=-N}^{N} y_o[n+i], \tag{3}$$

where $2N+1$ corresponds to the number of samples in the window used for trend removal.

The ZFF signal is used to derive the glottal source characteristics [73]. The positive-to-negative zero-crossings (PNZCs) correspond to GCIs (or epochs) by considering the negative polarity of the signal [73], [88]. Let us denote epochs by $\mathcal{E} = \{e_1, e_2, ..., e_M\}$, where M is the number of epochs. The time duration between any two adjacent epochs gives the instantaneous fundamental period ($T_0[k]$), and its reciprocal

gives the instantaneous fundamental frequency ($F_0[k]$), i.e.,

$$T_0[k] = \frac{(e_k - e_{k-1})}{f_s}, \qquad k = 2, 3, ..., M, \qquad (4)$$

$$F_0[k] = \frac{1}{T_0[k]} = \frac{f_s}{(e_k - e_{k-1})}, \qquad k = 2, 3, ..., M, (5)$$

where f_s is the sampling frequency.

Another interesting property of the ZFF signal is that the slope of the signal around each PNZC is proportional to the rate of closure of the vocal folds as measured using differentiated electroglottography (EGG) signals at the instants of glottal closure. A block diagram describing the steps involved in the ZFF method is shown in Fig. 2.
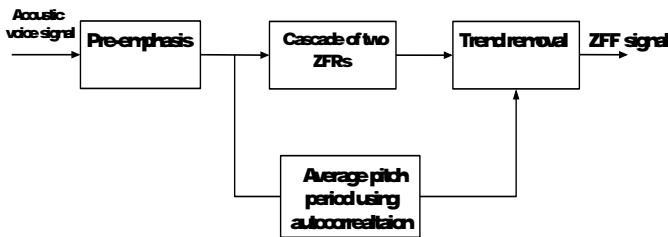

Fig. 2. Block diagram of the ZFF method.

To illustrate the glottal source waveforms computed by QCP and ZFF, a segment of voiced speech along with the simultaneously recorded EGG signal from the CMU ARCTIC database [89] is used. Fig. 3(a) and Fig. 3(b) show the acoustic speech signal and the differentiated EGG, respectively. Glottal source waveforms computed by QCP and ZFF are shown in Fig. 3(c) and Fig. 3(d), respectively.
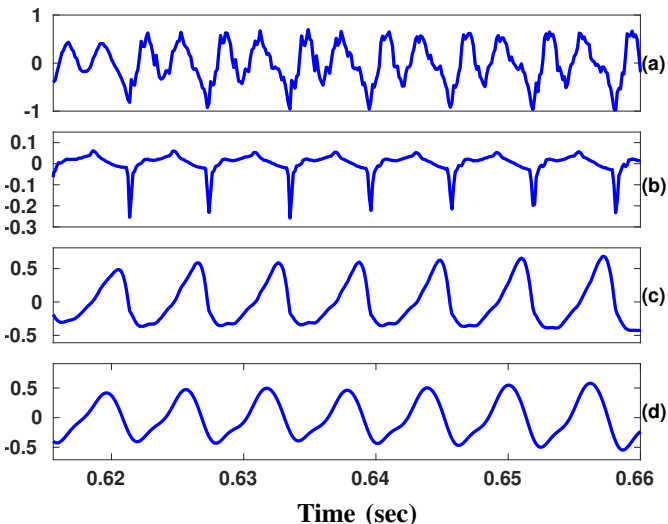

Fig. 3. Illustration of glottal source waveforms derived using the QCP and ZFF methods: (a) acoustic speech signal, (b) differentiated EGG signal, (c) glottal source waveform estimated by QCP, and (d) approximate glottal source waveform estimated by ZFF (reversed in polarity for visualization purpose).

III. EXTRACTION OF GLOTTAL SOURCE FEATURES

This section describes the extraction of features from the glottal source waveforms computed using QCP and ZFF. In addition, the section explains the extraction of the glottal source features that are computed directly from the acoustic voice signal and that capture specific properties of the glottal source.

A. Glottal source features derived using the QCP method

In order to represent the glottal flow waveform in a compact form, different methods have been developed for parameterization and they can be grouped into two categories: time-domain and frequency-domain glottal features (also called glottal parameters).

1) Time-domain glottal features: Time-domain glottal flow signals can be parameterized using time-based and amplitude-based features [74], [75]. In the case of time-based features, the most classical approach is to compute time-duration ratios between the different phases (closed phase, opening phase, and closing phase) of the glottal source waveform in a glottal cycle. The measures are defined by extracting critical time instants (such as the instant of glottal closure, primary and secondary glottal opening, the instant of minimum and maximum glottal flow) from the glottal source waveform. In the case of amplitude-based features (amplitude quotient [90], [91] and normalized amplitude quotient [92]), the amplitude of the glottal flow and its derivative are used [90], [92], [93]. The normalized amplitude quotient has been shown to be a strong correlate of the closing quotient, and it has been extensively used in analyzing voice quality [92]. Extraction of critical time instants is often difficult and to overcome this, sometimes time-based features are computed by replacing the true closure and opening instants by the time instants when the glottal flow crosses a level, which is set to a value between the maximum and minimum amplitude of glottal flow in a glottal cycle [75].

2) Frequency-domain glottal features: While the computation of time-domain features from the glottal source waveform is straightforward, these features are affected by distortions such as formant ripple due to incomplete canceling of formants by the inverse filter [75]. In such cases, it is useful to derive frequency-domain features for the glottal source waveform. Frequency-domain features are computed from the spectrum of the glottal source and they essentially measure the slope of the spectrum. Several studies have quantified the spectral slope of the glottal source by utilizing the level of F_0 and its harmonics. The most widely used features are the amplitude difference between F_0 and the first harmonic (H1-H2) [94], the harmonic richness factor (HRF) [95], and the parabolic spectral parameter (PSP) [96]. HRF is the ratio between the sum of the amplitudes of the harmonics above F_0 and the amplitude of F_0. PSP is derived by fitting a parabola to the low frequencies of the glottal flow spectrum [96].

A total of 12 glottal features (9 time-domain and 3 frequency-domain features) defined in [74] are used in this study to characterize the glottal flow waveforms estimated by the QCP glottal inverse filtering method. These features are extracted using the APARAT Toolbox [74] and they are listed in Table I.

TABLE I
TIME-DOMAIN AND FREQUENCY-DOMAIN GLOTTAL FEATURES DERIVED
FROM GLOTTAL FLOWS ESTIMATED BY QCP.

	Time-domain features
OQ1	Open quotient, calculated from the primary glottal opening
OQ2	Open quotient, calculated from the secondary glottal opening
NAQ	Normalized amplitude quotient
AQ	Amplitude quotient
ClQ	Closing quotient
OQa	Open quotient, derived from the LF model
QoQ	Quasi-open quotient
SQ1	Speed quotient, calculated from the primary glottal opening
SQ2	Speed quotient, calculated from the secondary glottal opening
	Frequency-domain features
H1-H2	Amplitude difference between the first two glottal harmonics
PSP	Parabolic spectral parameter
HRF	Harmonic richness factor

B. Glottal source features derived using the ZFF method

From the ZFF method, the following glottal source features are extracted: the strength of excitation (SoE), energy of excitation (EoE), loudness measure and ZFF signal energy. These features have been shown to be useful for discriminating phonation types and emotions in [76], [97]. The four ZFF-based parameters are computed as follows.

1) Strength of excitation (SoE): The slope of the ZFF signal around each PNZC corresponds to the SoE, which is proportional to the rate of closure of the vocal folds [97]. A measure of SoE around the GCI is given by

$$SoE = |y[e_k + 1] - y[e_k - 1]|, \qquad k = 1, 2, ..., M. \quad (6)$$

where $y[n]$ is the ZFF signal (Eq. 3).

2) Energy of excitation (EoE): The EoE feature is computed from the samples of the Hilbert envelope ($h_e[i]$) of the LP residual over a 1-ms region around each GCI. This feature, defined below in Eq. 7, has been shown to measure vocal effort [97].

$$EoE = \frac{1}{2K + 1} \sum_{i=-K}^{K} h_e^2[i], \quad (7)$$

where 2K+1 corresponds to the number of samples in the 1-ms window.

3) Loudness measure: The loudness measure captures the abruptness of glottal closure [97], and it is defined according to Eq. 8 as the ratio between the standard deviation (σ) and mean (μ) of the samples of the LP residual's Hilbert envelope in a 1-ms region around GCI.

$$Loudness\ measure = \frac{\sigma}{\mu}. \quad (8)$$

4) ZFF signal energy ($v_{zff}[n]$): The energy of the ZFF signal is given by

$$v_{zff}[n] = \frac{1}{L} \sum_{i=-L/2}^{L/2} y^2[n+i], \quad (9)$$

where $y[n]$ is the ZFF signal. The energy of the ZFF signal at GCI is used in this study.

The steps involved in the extraction of glottal source features from the ZFF method are shown in the schematic block diagram in Fig. 4.
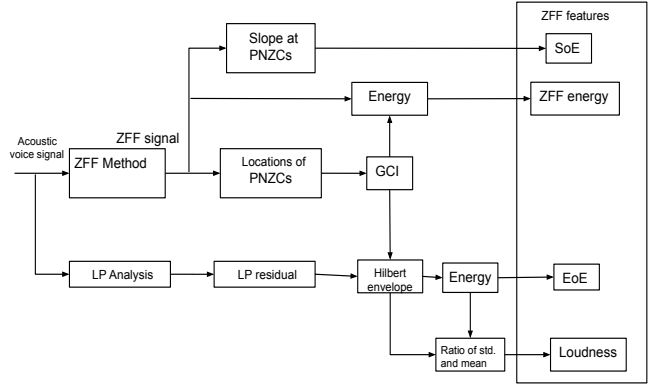

Fig. 4. Schematic block diagram for the extraction of glottal source features using the ZFF method.

C. Glottal source features derived directly from acoustic voice signals

The following four parameters, which capture the specific property of the glottal source, are computed directly from acoustic voice signals without computing the glottal source waveform.

1) Maximum dispersion quotient (MDQ): The MDQ parameter captures the abruptness of closure of the vocal folds [77]. This parameter measures the dispersion in the LP residual around GCI. Here, wavelet decomposition is carried out for the LP residual. Within a search interval near the GCI of the decomposed signals, the distance of the maxima locations to the given GCI is measured. The average of these distances normalized to the pitch period is referred to as MDQ.

2) Peak slope (PS): The PS parameter [78] captures the spectral slope of the glottal source. The method involves computing a wavelet-based decomposition of the acoustic voice signal into octave bands and then fitting a regression line to the maximum amplitudes at the different frequency bands. The slope coefficient of the fitted regression line is referred to as PS.

3) Cepstral peak prominence (CPP): The CPP parameter measures the amount of periodicity present in the signal using cepstrum [30]. A high CPP value reflects a more periodic structure in a signal. Initially this parameter was proposed to characterize the breathiness of voice signals [98]. CPP measures the difference between the most prominent cepstral peak (first rahmonic) and the point with the same quefrency on the regression line through the smoothed cepstrum.

4) Rd shape parameter: The Rd shape parameter is based on first presenting the entire glottal flow waveform using the parametric Liljencrants-Fant (LF) model [93]) and then presenting the LF pulse using a single parameter. The Rd shape parameter [79], [80] provides a single feature which captures most of the covariation of the LF parameters. A high value of Rd indicates a more relaxed voice.

IV. ANALYSIS OF NORMAL AND PATHOLOGICAL VOICE WITH GLOTTAL SOURCE FEATURES

This section presents results that were obtained when the glottal source features described in Section III were used to analyze normal and pathological voice. The analyses were carried out using the twenty speakers of the HUPA database (details of the database are given in Section VI-A1). The results obtained are described in feature distributions that are depicted using box plots. By presenting box plots of the feature distributions, our aim is to analyze potential differences between different glottal source features in their discriminability of normal and pathological voice.

A. Analysis of the glottal source features derived using the QCP method

Figure 5 shows distributions of the glottal source features derived using the QCP method for normal and pathological voice. The figure shows the nine time-domain features (rows 1, 2 and 3) and the three frequency-domain features (row 4). It can be seen that the frequency-domain features result in better discrimination of normal and pathological voice compared to the time-domain features. In the time-domain features, NAQ discriminates normal and pathological voice better than the other features. For the open quotients, OQ1 and OQ2 show larger variations in pathological voice compared to normal speech, and QoQ indicates less discriminability. On the other hand, the LF model-based open quotient (OQa) shows good discriminability. AQ, ClQ, SQ1 and SQ2 show in general small differences in distributions between normal and pathological voice. This may be due to the difficulty in identifying critical glottal time instants (instant of glottal closure, primary and secondary glottal opening).

The frequency-domain features show in general better discriminability of normal and pathological voice compared to the time-domain features. The values of H1-H2 and PSP are higher in pathological voice compared to normal indicating that the spectral slope of the glottal source is deeper. On the other hand, HRF values are lower for pathological voice due to a weaker harmonic structure (see Fig. 8) in their glottal source spectrum.

B. Analysis of the glottal source features derived using the ZFF method

Figure 6 shows the distribution of the glottal source features derived using the ZFF method. It can be seen that all the features show in general good discriminability of normal and pathological voice. SoE, which measures the strength of the impulse-like excitation at glottal closure, is lower in pathology
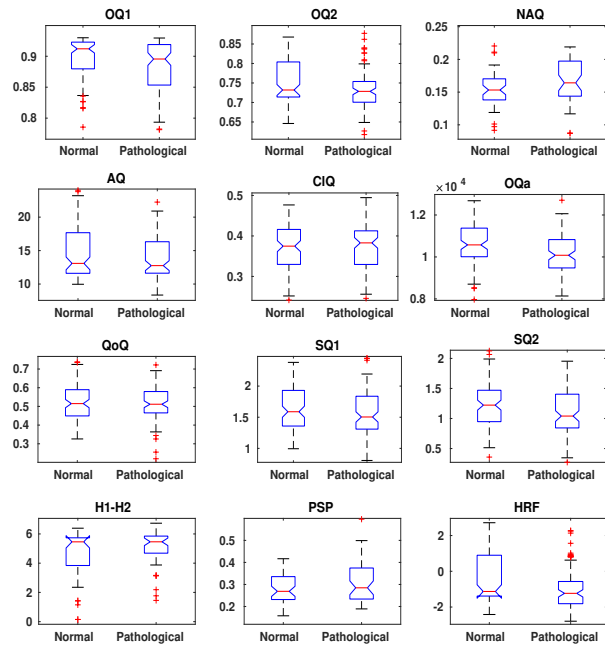

Fig. 5. Distribution of the glottal source features derived from the QCP method for normal and pathological voice using box plots. The central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers on either side cover all points within 1.5 times the interquartile range, and points beyond these whiskers are plotted as outliers using the $'+'$ symbol.

indicating less abrupt closure of the vocal folds compared to normal speech. EoE, which measures the energy of excitation at the glottal closure and captures the vocal effort required to produce the voice signal, is also lower in pathology compared to normal. As pathological voice is produced with improper and slower glottal closure, the loudness measure values are also lower. The ZFF signal energy of pathological voice is lower than in normal voice, similar to EoE.

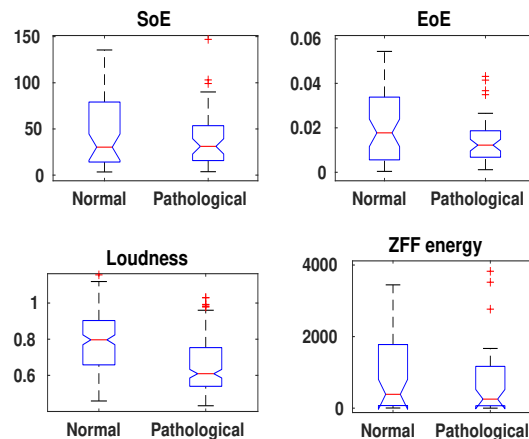Fig. 6. Distribution of the glottal source features derived from the ZFF method for normal and pathological voice.The central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers on either side cover all points within 1.5 times the interquartile range, and points beyond these whiskers are plotted as outliers using the $'+'$ symbol.

C. Analysis of the glottal source features derived from acoustic voice signals

Figure 7 shows the distribution of the glottal source features derived directly from acoustic voice signals. It can be seen that all the features are capable of discriminating normal and pathological voice. The MDQ feature, which measures the dispersion of the LP residual around glottal closure, is high in pathology indicating the occurrence of improper glottal closure and increased aspiration noise. PS, which captures the spectral slope of the glottal source, is also higher in pathology. This observation is not very evident from the figure, as the range of the PS values is higher in normal voice. As pathological voice is produced with a larger amount of aspiration noise and improper glottal closure, it is similar to normal voice of breathy phonation. For breathy or relaxed voices, the Rd feature values are high [79], [80]. This observation is evident from the box plot of the Rd feature. The CPP feature measures the amount of periodicity present in the signal. Because of improper glottal closure and an increased amount of aspiration noise in pathological voice, the harmonicity of the glottal source spectrum is weaker (see Fig. 8). Hence, the CPP values are lower for pathological voice, which is evident from the box plot.
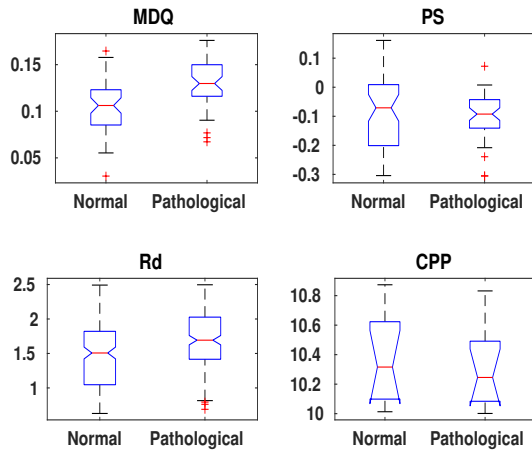

Fig. 7. Distribution of the glottal source features derived directly from acoustic speech signals for normal and pathological voice. The central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers on either side cover all points within 1.5 times the interquartile range, and points beyond these whiskers are plotted as outliers using the $'+'$ symbol.

V. EXTRACTION OF MFCCs FROM GLOTTAL SOURCE WAVEFORMS

From the analysis of the glottal source features described in Section IV-A, it can be concluded that the features derived from the glottal source spectrum (frequency-domain features) have a better discriminability compared to the time-domain features. This motivates us to use the entire glottal source spectrum, instead of a few single features, in voice pathology detection. Figure 8 shows spectrograms of glottal flow waveforms for normal and pathological voice estimated using the QCP method. It can be seen that there are large variations especially in the harmonic structure of the glottal flow spectra between normal and pathological voice. In order to capture

these variations and to represent them in a compact form, we propose to derive MFCCs from the spectra of the glottal source waveforms (as in our recent conference paper [99]). It should be noted that the proposed MFCC feature extraction is similar to the computation of conventional MFCC features, except that the proposed approach operates on the glottal source waveform instead of the acoustic voice signal.

A schematic block diagram of the extraction of MFCCs from the glottal source waveform given by QCP and ZFF is shown in Fig. 9. The method involves short-term spectral analysis, where the glottal source waveform is split into overlapping time-frames and the spectrum of each frame is computed with DFT. The spectrum is estimated using a 1024-point DFT with Hamming windowing in 25-ms frames with a 5-ms shift. Mel-cepstrum is derived from the mel-scale-based analysis of the spectrum of the glottal source, followed by logarithm and discrete cosine transform (DCT). From the entire mel-cepstrum, the first 13 coefficients (including the 0th coefficient) are considered for each frame. The resulting cepstral coefficients are referred as MFCC-QCP and MFCC-ZFF for the glottal source waveforms computed by QCP and ZFF, respectively. From static cepstral coefficients, delta and double-delta coefficients are also computed.
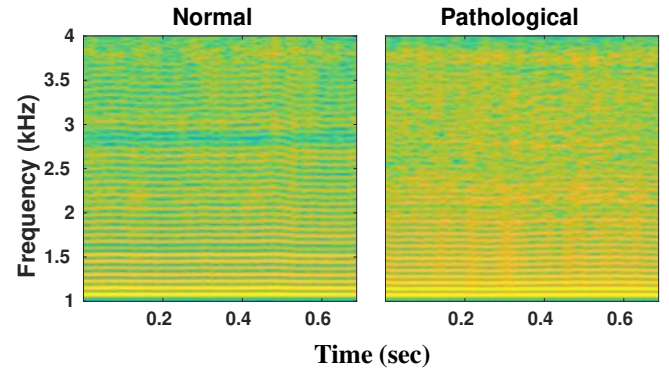

Fig. 8. Illustration of spectrograms of glottal source waveforms estimated using the QCP method for normal and pathological voice.

VI. EXPERIMENTAL PROTOCOL

This section describes the databases, the feature sets used in voice pathology detection including the baseline features, the classifier and the evaluation metrics.

A. Databases of pathological voice

Two databases containing normal and pathological voice are used in this study. These databases are the Hospital Universitario Príncipe de Asturias (HUPA) database [81], [82] and the Saarbrücken Voice Disorders (SVD) database [83], [84].

1) The HUPA database: This database was recorded at the Príncipe de Asturias hospital in Alcalá de Henares, Madrid, Spain [81], [82]. The dataset contains sustained phonations of the vowel /a/ by 439 adult Spanish speakers (239 healthy and 200 pathological). Originally, the data was recorded with a sampling frequency of 50 kHz and later downsampled to
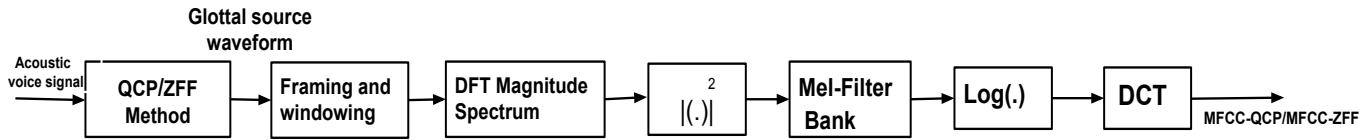
Fig. 9. Extraction of MFCCs from the glottal source waveforms computed by the QCP and ZFF methods.

25 kHz. Pathological voices contain a wide variety of organic pathologies such as nodules, polyps, oedemas and carcinomas. More details of the database can be found in [81], [82], [100].

2) The SVD database: This database was recorded at the Institut für Phonetik at Saarland University and the Phoniatry Section of the Caritas Clinic St. Theresia in Saarbrücken, Germany [83], [84]. The data comprises 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 dataset was recorded from 2225 German speakers, of which 869 are healthy and 1356 are pathological. The database contains 71 different pathologies including both functional and organic pathologies. The data was recorded with a sampling frequency of 50 kHz. As in [13], in this study we use the vowels /a/, /i/ and /u/, produced using normal pitch, and the running speech after removing the samples with a lower dynamic range, samples which are recorded after voice therapy and surgical intervention. This procedure resulted in data of 1518 speakers, of which 661 are healthy and 857 are pathological. More details of the database can be found in [83], [84].

B. The proposed glottal source feature sets and parameters used for feature extraction

In total, five sets of glottal source features are investigated as listed below:

- Time-domain (OQ1, OQ2, NAQ, ClQ, SQ1, SQ2, AQ, QoQ, OQa) and frequency-domain (H1-H2, PSP, HRF) features derived from the glottal source waveforms computed by the QCP method. These features are extracted for every glottal cycle and QCP analysis is carried out in Hamming-windowed 25-ms frames with a 5-ms shift.
- Features derived from the approximate glottal source signals computed by the ZFF method (SoE, EoE, loudness measure, ZFF energy). All these features are computed around GCIs. EoE and loudness measure are computed from the samples of the Hilbert envelope of the LP residual (computed with 12th order) over a 1-ms region around each GCI.
- Features that capture the specific property of the glottal source computed directly from acoustic voice signals without computing the glottal source waveform (MDQ, PS, CPP, Rd). All these features are computed in 25-ms Hamming-windowed frames with a 5-ms shift.
- MFCC-QCP is computed from the glottal flow waveforms estimated by QCP in 25-ms Hamming windowed frames with a 5-ms shift. First 13 static cepstral coefficients and their delta & double-delta coefficients are computed yielding a 39-dimensional feature vector.

- MFCC-ZFF is computed from the approximate glottal source waveforms given by ZFF using 25-ms Hamming windowed frames with a 5-ms shift. Here also, static coefficients and their delta & double-delta coefficients are computed yielding a 39-dimensional feature vector.

C. Baseline features used for comparison

We consider conventional MFCC and PLP features for comparison as they were shown in [13] to provide good discrimination between normal and pathological voice.

1) Mel-frequency cepstral coefficients (MFCCs): Conventional MFCC features were computed using 25-ms Hamming-windowed frames with a 5-ms shift. The first 13 cepstral coefficients (including the 0th coefficient) and their delta & double-delta coefficients were computed yielding a 39-dimensional feature vector.

2) Perceptual linear prediction (PLP) coefficients: Conventional PLP features were computed using 25-ms Hamming-windowed frames with a 5-ms shift. The first 13 cepstral coefficients (including the 0th coefficient) and their delta & double-delta coefficients were computed yielding a 39-dimensional feature vector.

D. Classifier

The most popular classifier for voice pathology detection is support vector machine (SVM). In the current study, we use SVM with a radial basis function (RBF) kernel. Experiments were conducted with 20-fold cross-validation, where the data was partitioned randomly into 20 equal portions. One fold was held out to be used for testing with the remaining nineteen folds for training. 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. Evaluation metrics were saved in each fold and this process was repeated for each of the 20 folds. Finally, the evaluation metrics were averaged over the 20 folds for evaluation.

E. Evaluation metrics

Standard performance metrics for a binary classifier are considered for each one of the aforementioned feature sets [13], [101]. Therefore, the following metrics are used: accuracy (ACC), sensitivity (SE), specificity (SP), area under the receiver operating characteristic curve (AUC), and equal error rate (EER). For a better performing system, the values of first four metrics should be higher and the last metric should be lower.

VII. PATHOLOGY DETECTION EXPERIMENTS

Pathology detection experiments were carried out using the SVM classifier with the individual feature sets described in Sections VI-B and VI-C as well as with combinations of the feature sets to analyze the complementary information among the features. In the combination of features, the complementary information among the glottal source feature sets and complementary information with the existing spectral features was also investigated. In total, 12 feature sets were investigated, out of which seven were individual feature sets (denoted FS-1 to FS-7) and five were combination of feature sets (denoted FS-8 to FS-12). These feature sets are listed below.

- FS-1: OQ1, OQ2, NAQ, ClQ, SQ1, SQ2, AQ, QoQ, OQa, H1-H2, PSP and HRF.
- FS-2: SoE, EoE, Loudness, ZFF energy.
- FS-3: MDQ, PS, CPP and Rd.
- FS-4: MFCC-QCP
- FS-5: MFCC-ZFF
- FS-6: Conventional MFCC features.
- FS-7: Conventional PLP features.
- FS-8: Combination of FS-1, FS-2 and FS-3
- FS-9: Combination of FS-4 and FS-5.
- FS-10: Combination of FS-1, FS-2, FS-3, FS-4 and FS-5 (combination of all glottal source features).
- FS-11: Combination of FS-6 and FS-7 (combination of spectral features i.e., MFCC and PLP).
- FS-12: Combination of FS-1, FS-2, FS-3, FS-4, FS-5, FS-6, and FS-7 (combination of all glottal source features and spectral features).

A total of five experiments were carried out: one experiment using the HUPA dataset (sustained phonation of the vowel /a/) and four experiments using the SVD dataset (sustained phonation of the vowels /a/, /i/, /u/, and the sentence sample), and the corresponding results are given in Tables II to VI.

Table II shows the voice pathology detection results computed for the vowel /a/ in the HUPA database with the individual feature sets (FS-1 to FS-7) and combination of feature sets (FS-8 to FS-12). From the table, it can be observed that in the case of individual feature sets, the feature set FS-5 (MFCC-ZFF) provided the best performance in terms of accuracy (72.89%), AUC (0.78) and EER (0.253). In terms of AUC and EER, the next best feature set was FS-4 (MFCC-QCP), which provided an AUC of 0.77 and EER of 0.267. The MFCC and PLP feature sets (FS-6 and FS-7) were also close to the MFCC-QCP features (FS-4). From the combination of feature sets (FS-8 to FS-12), it can be clearly seen that there exists an improvement in performance for all the combinations. This indicates the complementary information among the feature sets. Further, it is observed that the combination of MFCC-ZFF and MFCC-QCP (FS-9), and the combination of all glottal source features (FS-10) gave the highest detection performance. Also, it can be observed that the combination of conventional MFCC and PLP (FS-11) features showed an improvement in performance, which indicates the presence of complementary information between these features. Overall, the best performance was

observed when all glottal source feature sets (FS-1 to FS-5) and conventional MFCC and PLP features (FS-6 and FS-7) were combined. The combination of all the feature sets (FS-12) gave an accuracy of 78.37% , AUC of 0.84, and EER of 0.207, which highlights the complementary nature of the conventional features with the glottal source features for voice pathology detection.

TABLE II
RESULTS FOR THE VOWEL /A/ IN THE HUPA DATABASE WITH INDIVIDUAL FEATURE SETS (FS-1 TO FS-7) AND COMBINATION OF FEATURE SETS (FS-8 TO FS-12). SE REFERS TO SENSITIVITY, SP REFERS TO SPECIFICITY, AUC REFERS TO AREA UNDER THE ROC CURVE AND EER REFERS TO EQUAL ERROR RATE.

Feature set	Accuracy [%]	SE	SP	AUC	EER
FS-1	68.51±4.80	0.76	0.62	0.72	0.304
FS-2	66.89±3.54	0.71	0.63	0.71	0.315
FS-3	64.33±4.92	0.78	0.55	0.71	0.322
FS-4	70.28±4.14	0.75	0.66	0.77	0.267
FS-5	**72.89**±4.02	0.85	0.63	**0.78**	**0.253**
FS-6	72.00±4.36	0.80	0.63	0.73	0.279
FS-7	72.36±4.23	0.78	0.67	0.74	0.272
FS-8	73.74±4.12	0.74	0.73	0.78	0.249
FS-9	76.67±5.01	0.81	0.72	**0.79**	0.241
FS-10	**76.96**±4.12	0.80	0.74	**0.79**	**0.238**
FS-11	73.68±4.22	0.78	0.69	0.76	0.249
FS-12	**78.37**±4.18	0.80	0.77	**0.84**	**0.207**

Tables III, IV and V show the detection results for the vowels /a/, /i/ and /u/ in the SVD database with the individual feature sets (FS-1 to FS-7) and combination of feature sets (FS-8 to FS-12). As in the results obtained for the HUPA database, here also (in the case of individual feature sets FS-1 to FS-7), the MFCC-ZFF features (FS-5) provided the best performance in terms of AUC and EER in all vowels. In terms of accuracy, the PLP features gave the highest performance for all the three vowels. Complementary information among the features can be observed from the results of feature sets from FS-8 to FS-12. The combination of all the glottal source features (FS-10) gave the highest AUC and lowest EER for all the three vowels. Overall, the best performance was achieved when all the glottal source feature sets were combined with the conventional MFCC and PLP features (i.e., FS-12). The combination of all the feature sets gave an AUC of 0.76, 0.75 and 0.73, EER of 0.271, 0.284, and 0.292, and accuracy of 74.32%, 72.29% and 71.45% for the vowels /a/, /i/ and /u/, respectively. It should be noted that the detection experiments with the vowel /a/ achieved better results (highest AUC, accuracy and lowest EER) compared to the vowels /i/ and /u/. This might be due to the use of wider vocal tract openings in the production of the vowel /a/ which results in less interaction between the glottal source and the vocal tract [13]. This observation is also in line with results reported in [13].

Table VI shows the results for the continuous speech sentences in the SVD database with individual feature sets (FS-1 to FS-7) and combination of feature sets (FS-8 to FS-

TABLE III
RESULTS FOR THE VOWEL /A/ IN THE SVD DATABASE WITH INDIVIDUAL FEATURE SETS (FS-1 TO FS-7) AND COMBINATION OF FEATURE SETS (FS-8 TO FS-12). SE REFERS TO SENSITIVITY, SP REFERS TO SPECIFICITY, AUC REFERS TO AREA UNDER THE ROC CURVE AND EER REFERS TO EQUAL ERROR RATE.

Feature set	Accuracy [%]	SE	SP	AUC	EER
FS-1	65.91±4.82	0.70	0.62	0.68	0.341
FS-2	66.76±4.19	0.72	0.62	0.70	0.332
FS-3	64.11±4.57	0.67	0.61	0.71	0.338
FS-4	65.12±4.28	0.68	0.63	0.72	0.329
FS-5	66.16±4.92	0.69	0.63	**0.74**	**0.307**
FS-6	67.25±4.11	0.70	0.64	0.72	0.315
FS-7	**68.19**±3.84	0.70	0.65	0.73	0.320
FS-8	68.64±4.02	0.69	0.66	0.73	0.319
FS-9	68.12±4.12	0.70	0.65	0.74	0.297
FS-10	**72.12**±4.35	0.73	0.69	**0.75**	**0.286**
FS-11	69.97±4.02	0.71	0.69	0.74	0.302
FS-12	**74.32**±3.58	0.75	0.71	**0.76**	**0.271**

TABLE IV
RESULTS FOR THE VOWEL /I/ IN THE SVD DATABASE WITH INDIVIDUAL FEATURE SETS (FS-1 TO FS-7) AND COMBINATION OF FEATURE SETS (FS-8 TO FS-12). SE REFERS TO SENSITIVITY, SP REFERS TO SPECIFICITY, AUC REFERS TO AREA UNDER THE ROC CURVE AND EER REFERS TO EQUAL ERROR RATE.

Feature set	Accuracy [%]	SE	SP	AUC	EER
FS-1	63.57±4.91	0.68	0.60	0.67	0.361
FS-2	65.12±4.73	0.70	0.60	0.67	0.353
FS-3	63.11±4.89	0.66	0.59	0.66	0.367
FS-4	65.51±4.48	0.69	0.61	0.67	0.359
FS-5	66.10±4.16	0.70	0.62	**0.70**	**0.322**
FS-6	66.81±4.22	0.71	0.61	**0.70**	0.323
FS-7	**67.79**±4.54	0.71	0.63	0.69	0.334
FS-8	67.04±4.63	0.69	0.64	0.70	0.326
FS-9	67.68±4.32	0.70	0.62	**0.71**	0.316
FS-10	66.91±4.27	0.69	0.63	**0.71**	**0.310**
FS-11	**69.11**±3.97	0.72	0.64	0.70	0.324
FS-12	**72.29**±4.02	0.74	0.67	**0.75**	**0.284**

TABLE V
RESULTS FOR THE VOWEL /U/ IN THE SVD DATABASE WITH INDIVIDUAL FEATURE SETS (FS-1 TO FS-7) AND COMBINATION OF FEATURE SETS (FS-8 TO FS-12). SE REFERS TO SENSITIVITY, SP REFERS TO SPECIFICITY, AUC REFERS TO AREA UNDER THE ROC CURVE AND EER REFERS TO EQUAL ERROR RATE.

Feature set	Accuracy [%]	SE	SP	AUC	EER
FS-1	63.12±4.22	0.65	0.61	0.66	0.361
FS-2	63.59±4.35	0.64	0.63	0.66	0.365
FS-3	62.11±4.08	0.62	0.61	0.64	0.378
FS-4	65.91±3.94	0.69	0.62	0.67	0.345
FS-5	65.77±4.19	0.70	0.61	**0.68**	**0.336**
FS-6	65.49±4.32	0.69	0.60	0.67	0.349
FS-7	**67.14**±3.92	0.71	0.62	**0.68**	0.343
FS-8	64.98±3.96	0.69	0.61	0.68	0.347
FS-9	66.89±3.48	0.70	0.63	0.68	0.330
FS-10	**68.84**±3.67	0.71	0.65	**0.71**	**0.319**
FS-11	68.22±3.84	0.70	0.66	0.69	0.328
FS-12	**71.45**±3.79	0.72	0.69	**0.73**	**0.292**

TABLE VI
RESULTS FOR THE SENTENCES IN THE SVD DATABASE WITH INDIVIDUAL FEATURE SETS (FS-1 TO FS-7) AND COMBINATION OF FEATURE SETS (FS-8 TO FS-12). SE REFERS TO SENSITIVITY, SP REFERS TO SPECIFICITY, AUC REFERS TO AREA UNDER THE ROC CURVE AND EER REFERS TO EQUAL ERROR RATE.

Feature set	Accuracy [%]	SE	SP	AUC	EER
FS-1	68.14±4.18	0.68	0.67	0.67	0.332
FS-2	65.28±4.35	0.69	0.61	0.65	0.343
FS-3	64.47±4.05	0.62	0.66	0.65	0.345
FS-4	65.95±3.92	0.64	0.67	0.68	0.327
FS-5	**68.92**±3.84	0.66	0.71	**0.70**	**0.318**
FS-6	67.89±4.20	0.65	0.69	0.68	0.331
FS-7	67.52±4.58	0.65	0.68	0.70	0.322
FS-8	70.81±3.81	0.71	0.70	**0.74**	**0.286**
FS-9	72.15±3.98	0.69	0.73	0.73	0.307
FS-10	**73.72**±4.03	0.69	0.75	0.73	0.296
FS-11	69.93±3.79	0.68	0.70	0.72	0.309
FS-12	**76.19**±3.62	0.72	0.78	**0.78**	**0.262**

12). In the case of individual feature sets, MFCC-ZFF (FS-5) achieved the highest AUC of 0.7 and lowest EER of 0.318. Conventional MFCCs (FS-6), the proposed MFCC-QCP (FS-4) and MFCC-ZFF (FS-5) had nearly similar performance. The results of the combination of feature sets (FS-8 to FS-12) indicate the complementary nature of the feature sets. In the case of combination of feature sets, 1-dimensional glottal source features (combination of QCP features, ZFF features and features derived directly from voice signals) gave the highest AUC of 0.74 and lowest EER of 0.286. Overall, the best performance was achieved (EER of 0.262, AUC of 0.78 and accuracy of 76.19%) when all the feature sets were combined, indicating the complementary nature of information of the glottal source features with the existing conventional

spectral features, MFCCs and PLPs.

It is worth noting that there exist studies in the literature [4], [67], [82] which report detection performance superior to that obtained in this study, but many of those studies have only included a small portion of the database and/or limited the analyses to a restricted number of pathologies. It is observed that the trend in the results reported in this paper are in line with the results reported in [13], [85].

VIII. SUMMARY

Glottal source features were studied in this work in the analysis of normal and pathological voice, and these features were further used in voice pathology detection. The glottal source features were derived from three signals: from the

glottal flows estimated with the QCP inverse filtering method, from the approximate source signals computed with the ZFF method and directly from acoustic voice signals. Analysis of features revealed that glottal source features help in discriminating normal voice from pathological voice. Detection experiments were carried out using two databases with individual glottal source feature sets and with a combination of features. Experiments showed that on their own the studied glottal source features provide better discrimination compared to spectral features such as MFCCs and PLPs features. Also, it was shown that complementary information exists among the different glottal source features. Further, the combination of the existing spectral features with the glottal source features resulted in improved detection performance, indicating the complementary nature of features.

Motivated by the voice pathology detection performance achieved using glottal source features, we intend to use these features in the future for the classification of pathologies and for predicting the level of pathology (i.e., quantifying the severity level, for example, as mild, medium, high and very high), which may be helpful for diagnosis.

REFERENCES

[1] A. Aronson, *Clinical Voice Disorders; An Interdisciplinary Approach*. Thieme Inc, 1985.

[2] N. R. Williams, "Occupational groups at risk of voice disorders: a review of the literature," *Occupational Medicine*, vol. 53, no. 7, pp. 456–460, 2003.

[3] 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 Transactions on Biomedical Engineering*, vol. 58, no. 2, pp. 370–379, 2011.

[4] J. A. G. García, L. Moro-Velázquez, and J. I. Godino-Llorente, "On the design of automatic voice condition analysis systems. part I: review of concepts and an insight to the state of the art," *Biomedical Signal Processing and Control*, vol. 51, pp. 181 – 199, 2019.

[5] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, and L. O. Ramig, "Suitability of dysphonia measurements for telemonitoring of parkinson's disease," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 4, p. 1015, 2009.

[6] D. G. Silva, L. C. Oliveira, and M. Andrea, "Jitter estimation algorithms for detection of pathological voices," *EURASIP Journal on Advances in Signal Processing*, pp. 9:1–9:9, Jan. 2009.

[7] M. Vasilakis and Y. Stylianou, "Voice pathology detection based on short-term jitter estimations in running speech," *Folia Phoniatrica Logopedica*, vol. 61, no. 3, pp. 153–170, 2009.

[8] Y. Zhang, J. J. Jiang, L. Biazzo, and M. Jorgensen, "Perturbation and nonlinear dynamic analyses of voices from patients with unilateral laryngeal paralysis," *Journal of Voice*, vol. 19, no. 4, pp. 519 – 528, 2005.

[9] V. Parsa and D. G. Jamieson, "Acoustic discrimination of pathological voice," *Journal of Speech, Language, and Hearing Research*, vol. 44, no. 2, pp. 327–339, 2001.

[10] J. R. Orozco-Arroyave, E. A. Belalcazar-Bolaos, J. D. Arias-Londoo, J. F. Vargas-Bonilla, S. Skodda, J. Rusz, K. Daqrouq, F. Hnig, and E. Nth, "Characterization methods for the detection of multiple voice disorders: Neurological, functional, and laryngeal diseases," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 6, pp. 1820–1828, Nov 2015.

[11] J. Mekyska, E. Janousova, P. Gomez-Vilda, Z. Smekal, I. Rektorova, I. Eliasova, M. Kostalova, M. Mrackova, J. B. Alonso-Hernandez, M. Faundez-Zanuy, and K. L. de Ipia, "Robust and complex approach of pathological speech signal analysis," *Neurocomputing*, vol. 167, pp. 94 – 111, 2015.

[12] N. Roy, J. Barkmeier-Kraemer, T. Eadie, M. P. Sivasankar, D. Mehta, D. Paul, and R. Hillman, "Evidence-based clinical voice assessment: A systematic review," *American Journal of Speech-Language Pathology*, 2013.

[13] J. A. G. 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," *Biomedical Signal Processing and Control*, vol. 48, pp. 128–143, 2019.

[14] C. Manfredi, M. D'Aniello, P. Bruscaglioni, and A. Ismaelli, "A comparative analysis of fundamental frequency estimation methods with application to pathological voices," *Medical Engineering & Physics*, vol. 22, no. 2, pp. 135 – 147, 2000.

[15] C. R. Watts and S. N. Awan, "Use of spectral/cepstral analyses for differentiating normal from hypofunctional voices in sustained vowel and continuous speech contexts," *Journal of Speech, Language, and Hearing Research*, vol. 54, no. 6, pp. 1525–1537, 2011.

[16] R. R. Patel, S. N. Awan, J. Barkmeier-Kraemer, M. Courey, D. Deliyski, T. Eadie, D. Paul, J. G. Švec, and R. Hillman, "Recommended protocols for instrumental assessment of voice: American speech-language-hearing association expert panel to develop a protocol for instrumental assessment of vocal function," *American Journal of Speech-Language Pathology*, vol. 27, no. 3, pp. 887–905, 2018.

[17] Y. Qi and R. E. Hillman, "Temporal and spectral estimations of harmonics-to-noise ratio in human voice signals," *The Journal of the Acoustical Society of America*, vol. 102, no. 1, pp. 537–543, 1997.

[18] J. Lee, S. Kim, and H. Kang, "Detecting pathological speech using contour modeling of harmonic-to-noise ratio," in *ICASSP*, May 2014, pp. 5969–5973.

[19] H. Kasuya, S. Ogawa, K. Mashima, and S. Ebihara, "Normalized noise energy as an acoustic measure to evaluate pathologic voice," *The Journal of the Acoustical Society of America*, vol. 80, no. 5, pp. 1329–1334, 1986.

[20] J. I. Godino-Llorente, V. Osma-Ruiz, N. Sáenz-Lechón, P. Gómez-Vilda, M. Blanco-Velasco, and F. Cruz-Roldán, "The effectiveness of the glottal to noise excitation ratio for the screening of voice disorders," *Journal of Voice*, vol. 24, no. 1, pp. 47–56, 2010.

[21] V. Parsa and D. G. Jamieson, "Identification of pathological voices using glottal noise measures," *Journal of Speech, Language, and Hearing Research*, vol. 43, no. 2, pp. 469–485, 2000.

[22] D. Michaelis, T. Gramss, and H. W. Strube, "Glottal-to-noise excitation ratio a new measure for describing pathological voices," *Acta Acustica united with Acustica*, vol. 83, no. 4, pp. 700–706, 1997.

[23] R. Fraile, J. I. Godino-Llorente, N. Sáenz-Lechón, J. M. Gutiérrez-Arriola, and V. Osma-Ruiz, "Spectral analysis of pathological voices: sustained vowels vs running speech," in *MAVEBA*, 2011, pp. 67–70.

[24] A. Benba, A. Jilbab, and A. Hammouch, "Discriminating between patients with parkinsons and neurological diseases using cepstral analysis," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 24, no. 10, pp. 1100–1108, Oct 2016.

[25] J. I. Godino-Llorente and P. G. Vilda, "Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 2, pp. 380–384, 2004.

[26] J. I. Godino-Llorente, S. Aguilera-Navarro, and P. G. Vilda, "Lpc, LPCC and MFCC parameterisation applied to the detection of voice impairments," in *INTERSPEECH*, 2000, pp. 965–968.

[27] J. C. Saldanha, T. Ananthakrishna, and R. Pinto, "Vocal fold pathology assessment using mel-frequency cepstral coefficients and linear predictive cepstral coefficients features," *Journal of Medical Imaging and Health Informatics*, vol. 4, no. 2, pp. 168–173, 2014.

[28] M. A. Little, D. A. Costello, and M. L. Harries, "Objective dysphonia quantification in vocal fold paralysis: Comparing nonlinear with classical measures," *Journal of Voice*, vol. 25, no. 1, pp. 21 – 31, 2011.

[29] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[30] J. Hillenbrand, R. A. Cleveland, and R. L. Erickson, "Acoustic correlates of breathy vocal quality," *Journal of Speech, Language, and Hearing Research*, vol. 37, no. 4, pp. 769–778, 1994.

[31] R. Fraile, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, and P. G. Vilda, "Use of cepstrum-based parameters for automatic pathology detection on speech - analysis of performance and theoretical justification," in *BIOSIGNALS*, 2008, pp. 85–91.

[32] R. Fraile and J. I. Godino-Llorente, "Cepstral peak prominence: A comprehensive analysis," *Biomedical Signal Processing and Control*, vol. 14, pp. 42–54, 2014.

[33] T. Drugman, T. Dubuisson, and T. Dutoit, "On the mutual information between source and filter contributions for voice pathology detection," in *INTERSPEECH*, 2009.

[34] R. Fraile, J. I. Godino-Llorente, N. Senz-Lechn, V. Osma-Ruiz, and J. M. Gutirrez-Arriola, "Characterization of dysphonic voices by means of a filterbank-based spectral analysis: Sustained vowels and running speech," *Journal of Voice*, vol. 27, no. 1, pp. 11 – 23, 2013.

[35] K. Umapathy, S. Krishnan, V. Parsa, and D. G. Jamieson, "Discrimination of pathological voices using a time-frequency approach," *IEEE Transactions on Biomedical Engineering*, vol. 52, no. 3, pp. 421–430, March 2005.

[36] B. Ghoraani and S. Krishnan, "A joint time-frequency and matrix decomposition feature extraction methodology for pathological voice classification," *EURASIP Journal on Advances in Signal Processing*, no. 1, Sep 2009.

[37] R. Behroozmand and F. Almasganj, "Optimal selection of wavelet-packet-based features using genetic algorithm in pathological assessment of patients speech signal with unilateral vocal fold paralysis," *Computers in Biology and Medicine*, vol. 37, no. 4, pp. 474 – 485, 2007.

[38] C. D. P. Crovato and A. Schuck, "The use of wavelet packet transform and artificial neural networks in analysis and classification of dysphonic voices," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 10, pp. 1898–1900, Oct 2007.

[39] E. S. Fonseca, R. C. Guido, P. R. Scalassara, C. D. Maciel, and J. C. Pereira, "Wavelet time-frequency analysis and least squares support vector machines for the identification of voice disorders," *Computers in Biology and Medicine*, vol. 37, no. 4, pp. 571–578, 2007.

[40] A. Akbari and M. K. Arjmandi, "An efficient voice pathology classification scheme based on applying multi-layer linear discriminant analysis to wavelet packet-based features," *Biomedical Signal Processing and Control*, vol. 10, pp. 209 – 223, 2014.

[41] M. Markaki and Y. Stylianou, "Using modulation spectra for voice pathology detection and classification," in *EMBC*, Sep. 2009, pp. 2514–2517.

[42] M. E. Markaki, Y. Stylianou, J. D. Arias-Londõno, and J. I. Godino-Llorente, "Dysphonia detection based on modulation spectral features and cepstral coefficients," in *ICASSP*, 2010, pp. 5162–5165.

[43] L. Moro-Velzquez, J. A. Gmez-Garca, and J. I. Godino-Llorente, "Voice pathology detection using modulation spectrum-optimized metrics," *Frontiers in Bioengineering and Biotechnology*, vol. 4, p. 1, 2016.

[44] M. Kaleem, B. Ghoraani, A. Guergachi, and S. Krishnan, "Pathological speech signal analysis and classification using empirical mode decomposition," *Medical & Biological Engineering & Computing*, vol. 51, no. 7, pp. 811–821, Jul 2013.

[45] M. A. Little, P. E. McSharry, S. J. Roberts, D. A. Costello, and I. M. Moroz, "Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection," *Biomedical engineering online*, vol. 6, no. 1, p. 23, 2007.

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

[47] P. Henríquez, J. B. Alonso, M. A. Ferrer, C. M. Travieso, J. I. Godino-Llorente, and F. Díaz-de-María, "Characterization of healthy and pathological voice through measures based on nonlinear dynamics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1186–1195, 2009.

[48] V. Péan, M. Ouayoun, C. Fugain, B. Meyer, and C. Chouard, "A fractal approach to normal and pathological voices," *Acta Otolaryngologica*, vol. 120, no. 2, pp. 222–224, 2000.

[49] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average parkinson's disease symptom severity," *Journal of The Royal Society Interface*, vol. 8, no. 59, pp. 842–855, 2011.

[50] G. Vaziri, F. Almasganj, and R. Behroozmand, "Pathological assessment of patients speech signals using nonlinear dynamical analysis," *Computers in Biology and Medicine*, vol. 40, no. 1, pp. 54 – 63, 2010.

[51] C. M. Travieso, J. B. Alonso, J. Orozco-Arroyave, J. Vargas-Bonilla, E. Nth, and A. G. Ravelo-Garca, "Detection of different voice diseases based on the nonlinear characterization of speech signals," *Expert Systems with Applications*, vol. 82, pp. 184 – 195, 2017.

[52] A. Giovanni, M. Ouaknine, and J.-M. Triglia, "Determination of largest lyapunov exponents of vocal signal: application to unilateral laryngeal paralysis," *Journal of Voice*, vol. 13, no. 3, pp. 341–354, 1999.

[53] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Accurate telemonitoring of parkinson's disease progression by noninvasive speech tests," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 4, pp. 884–893, April 2010.

[54] J. D. Arias-Londoo and J. I. Godino-Llorente, "Entropies from markov models as complexity measures of embedded attractors," *Entropy*, vol. 17, no. 6, pp. 3595–3620, 2015.

[55] J. A. G. García, J. I. Godino-Llorente, and G. Castellanos-Domínguez, "Influence of delay time on regularity estimation for voice pathology detection," in *EMBC*, 2012, pp. 4217–4220.

[56] M. Brockmann, M. J. Drinnan, C. Storck, and P. N. Carding, "Reliable jitter and shimmer measurements in voice clinics: The relevance of vowel, gender, vocal intensity, and fundamental frequency effects in a typical clinical task," *Journal of Voice*, vol. 25, no. 1, pp. 44 – 53, 2011.

[57] J. A. G. García, "Contributions to the design of automatic voice quality analysis systems using speech technologies," January 2018. [Online]. Available: http://oa.upm.es/49565/

[58] J. I. Godino-Llorente, P. G. 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 Transactions on Biomedical Engineering*, vol. 53, no. 10, pp. 1943–1953, 2006.

[59] S. Hadjitodorov, B. Boyanov, and B. Teston, "Laryngeal pathology detection by means of class-specific neural maps," *IEEE Transactions on Information Technology in Biomedicine*, vol. 4, no. 1, pp. 68–73, March 2000.

[60] M. K. Arjmandi, M. Pooyan, M. Mikaili, M. Vali, and A. Moqarehzadeh, "Identification of voice disorders using long-time features and support vector machine with different feature reduction methods," *Journal of Voice*, vol. 25, no. 6, pp. e275 – e289, 2011.

[61] W. Chen, C. Peng, X. Zhu, B. Wan, and D. Wei, "Svm-based identification of pathological voices," in *International Conference of the Engineering in Medicine and Biology Society*, 2007, pp. 3786–3789.

[62] J. D. Arias-Londoño, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, and G. Castellanos-Domínguez, "An improved method for voice pathology detection by means of a hmm-based feature space transformation," *Pattern Recognition*, vol. 43, no. 9, pp. 3100–3112, 2010.

[63] M. Alhussein and G. Muhammad, "Voice pathology detection using deep learning on mobile healthcare framework," *IEEE Access*, vol. 6, pp. 41 034–41 041, 2018.

[64] ——, "Automatic voice pathology monitoring using parallel deep models for smart healthcare," *IEEE Access*, vol. 7, pp. 46 474–46 479, 2019.

[65] G. Muhammad, S. M. M. Rahman, A. Alelaiwi, and A. Alamri, "Smart health solution integrating iot and cloud: A case study of voice pathology monitoring," *IEEE Communications Magazine*, vol. 55, no. 1, pp. 69–73, 2017.

[66] M. S. Hossain and G. Muhammad, "Healthcare big data voice pathology assessment framework," *iEEE Access*, vol. 4, pp. 7806–7815, 2016.

[67] S. Hegde, S. Shetty, S. Rai, and T. Dodderi, "A survey on machine learning approaches for automatic detection of voice disorders," *Journal of Voice*, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0892199718301437

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

[69] P. G. Vilda, R. Fernández-Baíllo, M. V. R. Biarge, V. N. Lluis, A. Á. Marquina, L. M. Mazaira-Fernández, R. Martínez-Olalla, and J. I. Godino-Llorente, "Glottal source biometrical signature for voice pathology detection," *Speech Communication*, vol. 51, no. 9, pp. 759–781, 2009.

[70] P. G. Vilda, J. I. Godino-Llorente, A. Á. Marquina, R. Martínez-Olalla, V. N. Lluis, and M. V. R. Biarge, "Evidence of glottal source spectral features found in vocal fold dynamics," in *ICASSP*, 2005, pp. 441–444.

[71] G. Muhammad, M. Alsulaiman, Z. Ali, T. A. Mesallam, M. Farahat, K. H. Malki, A. Al-nasheri, and M. A. Bencherif, "Voice pathology detection using interlaced derivative pattern on glottal source excitation," *Biomedical Signal Processing and Control*, vol. 31, pp. 156 – 164, 2017.

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

[73] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602–1613, Nov. 2008.

[74] M. Airas, "Tkk aparat: An environment for voice inverse filtering and parameterization," *Logopedics Phoniatrics Vocology*, vol. 33, no. 1, pp. 49–64, 2008.

[75] 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, 2011.

[76] S. R. Kadiri and B. Yegnanarayana, "Analysis and detection of phonation modes in singing voice using excitation source features and single frequency filtering cepstral coefficients (SFFCC)," in *INTERSPEECH*, 2018, pp. 441–445.

[77] J. Kane and C. Gobl, "Wavelet maxima dispersion for breathy to tense voice discrimination," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 6, pp. 1170–1179, 2013.

[78] ——, "Identifying regions of non-modal phonation using features of the wavelet transform," in *INTERSPEECH*, 2011, pp. 177–180.

[79] G. Degottex, A. Roebel, and X. Rodet, "Phase minimization for glottal model estimation," *IEEE Transactions on Acoustics, Speech and Language Processing*, vol. 19, no. 5, pp. 1080–1090, July 2011.

[80] S. Huber, A. Roebel, and G. Degottex, "Glottal source shape parameter estimation using phase minimization variants," in *INTERSPEECH*, Portland, USA, September 2012.

[81] L. Moro-Velázquez, J. Gómez-García, J. I. Godino-Llorente, and G. Andrade-Miranda, "Modulation spectra morphological parameters: A new method to assess voice pathologies according to the grbas scale," in *BioMed research international*, 2015.

[82] J. D. Arias-Londoo, J. I. Godino-Llorente, M. Markaki, and Y. Stylianou, "On combining information from modulation spectra and mel-frequency cepstral coefficients for automatic detection of pathological voices," *Logopedics Phoniatrics Vocology*, vol. 36, no. 2, pp. 60–69, 2011.

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

[84] ——, "Saarbrücken voice database, institute of phonetics, univ. of saarland," 2010, http://www.stimmdatenbank.coli.uni-saarland.de/ (Last viewed April 20, 2019).

[85] D. Martínez, E. Lleida, A. Ortega, A. Miguel, and J. Villalba, "Voice pathology detection on the saarbrücken voice database with calibration and fusion of scores using multifocal toolkit," in *Advances in Speech and Language Technologies for Iberian Languages*, 2012, pp. 99–109.

[86] D. Wong, J. Markel, and A. Gray, "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, pp. 350–355, 1979.

[87] P. Alku, J. Pohjalainen, M. Vainio, A.-M. Laukkanen, and B. H. Story, "Formant frequency estimation of high-pitched vowels using weighted linear prediction," *The Journal of the Acoustical Society of America*, vol. 134, no. 2, pp. 1295–1313, 2013.

[88] S. R. Kadiri and B. Yegnanarayana, "Speech polarity detection using strength of impulse-like excitation extracted from speech epochs," in *ICASSP*, 2017, pp. 5610–5614.

[89] CMU-ARCTIC speech synthesis databases. [Online]. Available: http://festvox.org/cmu_arctic/index.html

[90] P. Alku and E. Vilkman, "Amplitude domain quotient for characterization of the glottal volume velocity waveform estimated by inverse filtering," *Speech Communication*, vol. 18, pp. 131–138, 1996.

[91] P. Alku, M. Airas, E. Björkner, and J. Sundberg, "An amplitude quotient based method to analyze changes in the shape of the glottal pulse in the regulation of vocal intensity," *The Journal of the Acoustical Society of America*, vol. 120, pp. 1052–1062, 2006.

[92] P. Alku, T. Backstrom, and E. Vilkman, "Normalized amplitude quotient for parameterization of the glottal flow," *The Journal of the Acoustical Society of America*, vol. 112, pp. 701–710, 2002.

[93] G. Fant, "The lf-model revisited. transformations and frequency domain analysis," *Speech Transmission Laboratory Quarterly Progress and Status Report*, vol. 36, pp. 119–156, 1995.

[94] I. Titze and J. Sundberg, "Vocal intensity in speakers and singers," *The Journal of the Acoustical Society of America*, vol. 91, pp. 2936–2946, 1992.

[95] D. G. Childers and C. Lee, "Vocal quality factors: Analysis, synthesis, and perception," *The Journal of the Acoustical Society of America*, vol. 90, no. 5, pp. 2394–2410, 1991.

[96] P. Alku, H. Strik, and E. Vilkman, "Parabolic spectral parameter - A new method for quantification of the glottal flow," *Speech Communication*, vol. 22, no. 1, pp. 67–79, 1997.

[97] S. R. Kadiri, P. Gangamohan, S. V. Gangashetty, and B. Yegnanarayana, "Analysis of excitation source features of speech for emotion recognition," in *INTERSPEECH*, 2015, pp. 1324–1328.

[98] J. Hillenbrand and R. A. Houde, "Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech," *Journal of Speech, Language, and Hearing Research*, vol. 39, no. 2, pp. 311–321, 1996.

[99] S. R. Kadiri and P. Alku, "Mel-Frequency Cepstral Coefficients of Voice Source Waveforms for Classification of Phonation Types in Speech," in *Proc. Interspeech 2019*, 2019, pp. 2508–2512.

[100] 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," *European Archives of Oto-Rhino-Laryngology*, vol. 265, no. 4, pp. 465–476, Apr 2008.

[101] T. Fawcett, "An introduction to roc analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861 – 874, 2006.

Sudarsana Reddy Kadiri received Bachelor of Technology degree from Jawaharlal Nehru Technological University (JNTU), Hyderabad, India, in 2011, with a specialization in Electronics and Communication Engineering (ECE), the M.S. (Research) during 2011-2014, and later converted to Ph.D., and received Ph.D. degree from the Department of ECE, International Institute of Information Technology, Hyderabad (IIIT-H), India in 2018. He was a Teaching Assistant for several courses at IIIT-H during 2012-2018. He is currently a Postdoctoral Researcher with the Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland. His research interests include signal processing, speech analysis, speech synthesis, paralinguistics, affective computing, voice pathologies, machine learning and auditory neuroscience.

Paavo Alku received the M.Sc., Lic.Tech., and Dr.Sc.(Tech) degrees from the Helsinki University of Technology, Espoo, Finland, in 1986, 1988, and 1992, respectively. He was an Assistant Professor with the Asian Institute of Technology, Bangkok, Thailand, in 1993, and also with 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.