Kadiri, Sudarsana Reddy; Alku, Paavo

Glottal features for classification of phonation type from speech and neck surface accelerometer signals

Published in:
Computer Speech and Language

DOI:
10.1016/j.csl.2021.101232

Published: 01/11/2021

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

Computer Speech & Language 70 (2021) 101232

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

Glottal features for classification of phonation type from speech and neck surface accelerometer signals

Sudarsana Reddy Kadiri*, Paavo Alku

Department of Signal Processing and Acoustics, Aalto University, Finland

ARTICLE INFO

Article History:
Received 30 August 2020
Revised 13 April 2021
Accepted 16 April 2021
Available online 27 April 2021

Keywords:
Phonation type
Voice quality
Neck surface accelerometer
Glottal source waveform
Support vector machine

ABSTRACT

Glottal source characteristics vary between phonation types due to the tension of laryngeal muscles with the respiratory effort. Previous studies in the classification of phonation type have mainly used speech signals recorded by microphone. Recently, two studies were published in the classification of phonation type using neck surface accelerometer (NSA) signals. However, there are no previous studies comparing the use of the acoustic speech signal vs. the NSA signal as input in classifying phonation type. Therefore, the current study investigates simultaneously recorded speech and NSA signals in the classification of three phonation types (breathy, modal, pressed). The general goal is to understand which of the two signals (speech vs. NSA) is more effective in the classification task. We hypothesize that by using the same feature set for both signals, classification accuracy is higher for the NSA signal, which is more closely related to the physical vibration of the vocal folds and less affected by the vocal tract compared to the acoustical speech signal. Glottal source waveforms were computed using two signal processing methods, quasi-closed phase (QCP) glottal inverse filtering and zero frequency filtering (ZFF), and a group of time-domain and frequency-domain scalar features were computed from the obtained waveforms. In addition, the study investigated the use of mel-frequency cepstral coefficients (MFCCs) derived from the glottal source waveforms computed by QCP and ZFF. Classification experiments with support vector machine classifiers revealed that the NSA signal showed better discrimination of the phonation types compared to the speech signal when the same feature set was used. Furthermore, it was observed that the glottal features showed complementary information with the conventional MFCC features resulting in the best classification accuracy both for the NSA signal (86.9%) and the speech signal (80.6%).

1. Introduction

In producing speech signals, humans are capable of generating different phonation types (such as breathy, tense, creaky and falsetto) by regulating the activation of laryngeal muscles with the respiratory effort (Laver, 1980; Childers and Lee, 1991; Pietrowicz et al., 2017). Phonation type is closely associated with voice quality, a perceptual attribute defined as the auditory coloring of a speaker's voice (Laver, 1980). Breathy and tense/pressed voices are often considered to be the two opposite ends of the voice quality continuum (Kane and Gobl, 2013; Airas and Alku, 2007). Phonation type plays an important role in conveying para-linguistic information such as vocal emotions and personality in speech (Campbell and Mokhtari, 2003; Grichkovtsova et al., 2012;

*Corresponding author.
E-mail addresses: sudarsana.kadiri@aalto.fi (S.R. Kadiri), paavo.alku@aalto.fi (P. Alku).

https://doi.org/10.1016/j.csl.2021.101232

Park et al., 2018; Afshan et al., 2018; Birkholz et al., 2015). Breathy phonation is typically used in expressing politeness and intimacy (Ito, 2004), and tense phonation is used in expressing emotions of high activity such as anger, disgust, anxiety and excitement (Yanushevskaya et al., 2005; Gobl and Ní Chasaide, 2003). It is to be noted that apart from conveying para-linguistic information, phonation type is also used to generate phonological contrasts in certain languages (Gordon and Ladefoged, 2001; Ladefoged et al., 1988; Kuang and Keating, 2014; Esposito, 2010; ud Dowla Khan, 2012).

Modal phonation is typically used as the reference in comparing phonation types (Titze, 2000; Laver, 1980; Gordon and Ladefoged, 2001). In modal voices, the laryngeal tension settings are low and the vibration of the vocal folds is mostly periodic in a sequence of glottal cycles. In addition, the closure of the vocal folds is typically abrupt in modal voices. Breathy phonation, however, involves weaker levels of laryngeal tension and partial closure of the glottis which leads to the generation of turbulent noise (Titze, 2000; Laver, 1980). Hence, the harmonic structure of breathy voices is more prominent at low frequencies compared to modal voices (Gobl and Ní Chasaide, 2003; Alku, 2011). On the other hand, the laryngeal settings of pressed voices involve an increase in the adductive and longitudinal tension and a sharper glottal closure, which result in more prominent high-frequency harmonics compared to modal voices (Laver, 1980; Kane and Gobl, 2013; Titze, 2000; Alku, 2011). In the current study, the classification of phonation type into these three classes (breathy, modal and pressed) will be investigated. Literature review of the topic is given next by dividing the review into two parts based on the information signal that is used as input in the classification.

1.1. Classification of phonation type using speech signals

Variations in vibration patterns of the vocal folds result in differences in the shape of the glottal volume velocity pulse, called shortly as the glottal pulse, between phonation types. The glottal pulse varies from a smooth symmetric form in breathy phonation to an asymmetric form with sharp edges in pressed phonation as shown by studies where glottal inverse filtering (GIF) has been used to estimate the glottal pulse from speech (Airas and Alku, 2007; Alku et al., 2002b). This variation in the time-domain is reflected by the tilt of the glottal flow spectrum in the frequency-domain (Gowda and Kurimo, 2013; Hillenbrand et al., 1994). Using both time-domain and frequency-domain approaches, several features have been developed to parameterize and classify phonation type using the estimated glottal pulse waveforms (Airas and Alku, 2007; Kane and Gobl, 2013; Borsky et al., 2017b). Time-domain features such as the open quotient, the quasi-open quotient (QOQ) and the closing quotient (CQ), and amplitude-based features such as the amplitude quotient and the normalized amplitude quotient (NAQ) have been developed to parameterize the glottal pulse and its derivative (Airas and Alku, 2007; Alku, 2011; Drugman et al., 2014). Frequency-domain features such as the H1-H2 (the amplitude difference between the first (F0) and second harmonic) (Hillenbrand et al., 1994), the harmonic richness factor (HRF) (Childers and Lee, 1991) and the parabolic spectral parameter (PSP) (Alku et al., 1997) have been developed to measure the spectral tilt of the glottal pulse waveform. In Gobl and Ní Chasaide (2003); Swerts and Veldhuis (2001), phonation types were analysed by fitting the estimated glottal pulse derivative with an artificial Liljencrants-Fant (LF) glottal source model.

The performance of GIF deteriorates for high-pitched speech and expressive voices, which makes it difficult to derive glottal features, specially in the time-domain (Alku, 2011; Drugman et al., 2014). Due to the difficulties in applying GIF for high-pitched speech, studies in Garellek et al. (2016), Kreiman et al. (2012), Park et al. (2018) and Kreiman et al. (2015) measured the impact of the glottal source directly from the speech spectrum using features such as the F0, spectral slope between H4 (the fourth harmonic) and 2 kHz, and the spectral slope between 2 kHz and 5 kHz, H1-H2, and H2-H4. In Hillenbrand et al. (1994) and Klatt and Klatt (1990), cepstral peak prominence (CPP) was analyzed to capture the amount of aspiration noise in breathy phonation compared to modal phonation. Since breathy voices have larger open quotients and pressed voices have smaller open quotients, a spectral feature called the low-frequency spectral density (LFSD) was proposed in Gowda and Kurimo (2013). Due to the larger open quotient, the coupling of the subglottal system with the supraglottal system is stronger, making the low-frequency spectral energy larger in breathy voices compared to pressed voices.

The linear prediction (LP) residual and the zero frequency filtered signal (ZFFS) have also been used to derive features to classify phonation type. Features including the ZFFS slope, the energy of excitation, the loudness measure, and the ZFFS energy were used in the analysis and classification of phonation types in speech and singing (Kadiri and Yegnanarayana, 2018b; Kadiri et al., 2020). Their study showed that the ZFFS slope is inversely proportional to the duration of the glottal closed phase and that the ZFFS energy is directly proportional to the amount of low-frequency energy in speech signals. In addition, Kadiri et al. (2020) showed that the energy of excitation is directly proportional to the vocal effort, and that the loudness measure is directly proportional to the sharpness of glottal closure. In Kane and Gobl (2013), sharp changes in the glottal closure characteristics were captured using a measure called the maximum dispersion quotient (MDQ), which is based on the LP residual signal. It was reported that the discrimination capabilities of LFSD and MDQ were closer to the discrimination capability of NAQ (Gowda and Kurimo, 2013). Even though the harmonic-to-noise ratio (HNR) was found to provide poorer discrimination of phonation types, HNR has the potential to discriminate modal and breathy voices better compared to modal and pressed voices. In Kane and Gobl (2013) and Borsky et al. (2017b), a set of glottal source features along with the MFCCs derived from speech signals were investigated for the classification of phonation type. Glottal source features were also found to be useful in the detection of pathological voices in Narendra and Alku (2020). Recently, MFCCs derived from the zero-time windowing spectrum were explored for the classification of phonation type in singing in Kadiri and Alku (2019b). Moreover, a residual attention based neural network was studied in Sun et al. (2020) in the classification of phonation type in singing using the mel-spectrogram as input.

1.2. Classification of phonation type using neck surface accelerometer signals

The far majority of the phonation type classification studies is based on features which are computed from the acoustic speech signal by either estimating the glottal pulse with GIF or by using the speech spectrum directly. The neck surface accelerometer (NSA), however, provides an alternative means to indirectly measure vocal fold aerodynamics (Stevens et al., 1975; Rendon et al., 2007; Mehta et al., 2015). The accelerometer is a sensor which measures the vibration of the vocal folds in the direction normal to the neck surface during speech production. The accelerometer mounted on the below-glottis skin surface is less affected by the vocal tract. Therefore, this sensor can be considered to be free from the effect of formants which are the source of essential acoustic cues on speech intelligibility. The absence of formant cues makes the NSA signal unintelligible and therefore suitable for the real-life voice quality assessment and monitoring as it does not breach speaker privacy (Mehta et al., 2015; 2016).

NSAs have been used to measure laryngeal phenomena and it has been shown that NSA signals can supplement or supplant acoustic speech signals recorded by microphone (Coleman, 1988; Mehta et al., 2016; Titze et al., 2003). NSAs offer advantages over microphones by being much less sensitive to environmental noise sources. Moreover, when placed below the larynx, the NSA is also insensitive to articulatory modulation, thus protecting privacy (Mehta et al., 2016; Cortés et al., 2018). In addition, NSAs are conducive for use in ambulatory monitoring that are worn throughout an individual's daily activities to capture typical and atypical vocal behavior (Van Stan et al., 2015; Mehta et al., 2015). In addition to sensing phonation variations, accelerometers have also been useful to monitor subglottal resonances (Lulich et al., 2012) and nasal resonances (Stevens et al., 1975).

Relationships between acoustic-based estimates of the vocal function and their NSA-based counterparts have been investigated to gain an insight into the interpretability of accelerometer-based vocal features (Coleman, 1988; Titze and Hunter, 2015; Van Stan et al., 2015; Cortés et al., 2018). Strong correlations were found between speech and NSA signals in production of vowels for F0, jitter, CPP and HNR, and weak correlations were found for shimmer and spectral tilt (Mehta et al., 2016). The NSA signal was shown to accurately convey glottal features such as F0, maximum flow declination rate (MFDR, minimum level of the glottal flow derivative in one glottal cycle) and HRF which were derived using subglottal inverse filtering (Zañartu et al., 2013; Mehta et al., 2016; Lin et al., 2020). Recently, the relationship between the H1-H2 values computed from the NSA signal and from the glottal airflow waveform obtained by inverse filtering the oral airflow was studied (Mehta et al., 2019). It was found that the correlation between the two signals was high indicating a close relationship between them in glottal closure properties and skewness of the glottal source. In Ghassemi et al. (2014), it was shown that NSA signals have the capability of discriminating vocal hyperfunction from healthy production of speech using F0 and the sound pressure level, and their statistics. Several glottal features (such as the difference between the maximum and minimum amplitude within each glottal cycle, MFDR, open quotient, speed quotient, and NAQ) estimated from NSA signals were investigated for the assessment of vocal hyperfunction in Cortés et al. (2018).

Even though speech production has been studied using NSAs in several investigations as described above, there exist only two studies in the classification of phonation type using the NSA signal. In the first study (Borsky et al., 2017a), the authors investigated the discrimination capability of MFCCs derived from the NSA signal for four phonation types (modal, breathy, pressed, and rough voice). In Lei et al. (2019), features such as spectral harmonics, jitter, shimmer and spectral entropy extracted from the NSA signal were used for discriminating three phonation types (modal, breathy and pressed).

1.3. Goals of the study

The literature review in the two previous subsections indicates that the existing studies in the classification of phonation type focus on using features, which are extracted either from the glottal flow waveform (estimated by inverse filtering either the speech signal or the NSA signal) or from the spectrum of the speech signal or from the spectrum of the NSA signal. To the best of our knowledge, there are, however, no previous studies on the automatic classification of phonation type comparing features extracted from *simultaneous* recordings of the acoustic speech signal and the NSA signal. In other words, no one has yet studied which one of the two signals (the acoustic speech signal vs. the NSA signal) works better in the classification of phonation type. Therefore, the goal of the present study is to use different feature extraction methods to represent the speech signal and the (simultaneously recorded) NSA signal and to compare the automatic classification of breathy, modal and pressed vowels using the *same* set of features computed from the two inputs. We hypothesize that the classification accuracy should be higher when the NSA signal is used as input because the NSA signal carries information that is more directly related to the physiological vibratory patterns of the vocal folds compared the air flow generated by vocal fold vibration, the glottal pulse.

In summary, the highlights of the present study are:

- The classification of phonation types is studied for the first time by comparing simultaneously recorded speech and NSA signals.
- Three phonation types (breathy, modal, pressed) are studied.
- Glottal features are computed using the quasi-closed phase (QCP) glottal inverse filtering method and the zero frequency filtering (ZFF) method.
- Both scalar features and MFCCs derived from glottal waveforms are investigated.
- Experiments with the SVM classifier revealed that NSA signals discriminate phonation types better compared to speech signals.
- The combination of glottal and MFCC features showed improved classification accuracy both for speech and NSA signals.

1.4. Organization

The organization of the paper is as follows. Section 2 describes two signal processing methods, QCP and ZFF, for deriving glottal source waveforms. The extraction of one-dimensional glottal features and the extraction of MFCCs from the glottal source waveforms are described in Section 3. The experimental protocol is described in Section 4, which includes the database, the parameters used for feature extraction along with the feature sets used in the classification experiments and the details of the classifier. Results of the classification experiments are presented in Section 5. Finally, Section 6 summarizes the study.

2. Signal processing methods to compute glottal source waveforms

This section describes two signal processing methods which are used in the present study for the estimation of glottal source, the QCP glottal inverse filtering method (Airaksinen et al., 2014) and the ZFF method (Murty and Yegnanarayana, 2008). It is to be noted that inverse filtering of the acoustic speech signal recorded by microphone outside the mouth aims to remove the supraglottal resonances from the speech signal, whereas inverse filtering of the accelerometer signal aims to remove the subglottal resonances for deriving the glottal source.

Originally, the QCP and ZFF methods were proposed for acoustic speech signals. In the present study, we use these methods for processing both the acoustic speech signal as well as the NSA signal for extracting information about the glottal source. For the sake of simplicity of the presentation, the following descriptions of these signal processing methods are given using a generic time-domain input signal, denoted as $s[n]$, which refers both to the speech signal and the NSA signal.

2.1. The quasi-closed phase (QCP) method

QCP (Airaksinen et al., 2014), whose block diagram is shown in Fig. 1, is a GIF method which is based on the principles of closed phase analysis (Wong et al., 1979). Closed phase analysis is a GIF methods which estimates the vocal tract model from a few samples located in the closed phase of the glottal cycle using LP analysis. In contrast to closed phase analysis, the QCP method takes advantage of all the samples of the analysis frame in the computation of the vocal tract model. This is made possible by using weighted linear prediction (WLP) analysis, computed with the attenuated main excitation (AME) (Alku et al., 2013) weighting function, as an all-pole modeling method in the estimation of the vocal tract transfer function. The AME function is a straightforward time-domain waveform, using which WLP analysis can be made de-emphasize the square of the prediction error in those samples where the effect of the glottal excitation is prominent (i.e. in the vicinity of glottal closure). Consequently, the resulting all-pole WLP model (denoted by $V(z)$ in Fig. 1) is affected more by the characteristics of the vocal tract leading to smaller biasing of the vocal tract model by the glottal source. As shown in Alku et al. (2013), the AME weighting function is a simple, positive and real-valued waveform which is equal to 1.0 in the samples during the glottal open phase and equal to a small positive value (e.g. 0.03) in the vicinity of glottal closure. The AME waveform can be adjusted with three parameters (the duration quotient, the position quotient and the value of the waveform at glottal closure). In addition, the generation of the AME waveform calls for extracting glottal closure instants (GCIs). After computing the vocal tract model using WLP analysis with the AME weighting function, the input signal ($s[n]$) is finally inverse filtered in the QCP method with $V(z)$ to estimate the glottal source waveform. The QCP method was shown to provide better glottal source waveforms for modal and non-modal vowels compared to four existing inverse filtering methods (Airaksinen et al., 2014). Hence, in the current study, QCP is used as a GIF method to estimate the glottal source waveform.

2.2. The zero frequency filtering (ZFF) method

Based on the fact that the effect of an impulse-like excitation (which occurs at the instant of glottal closure) is present throughout the spectrum including the zero frequency (0 Hz), the ZFF method was proposed in Murty and Yegnanarayana (2008). In this method, the pre-emphasized signal ($x[n] = s[n] - s[n-1]$) is first passed through a cascade of two zero frequency resonators (ZFRs). That is, the pre-emphasized signal is filtered with a filter which has a pair of poles on the unit circle at the positive real axis in the z-plane and the filtering can be expressed as:

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

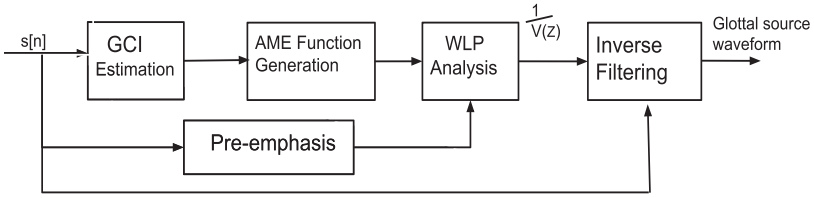

Fig. 1. Block diagram of the QCP method.

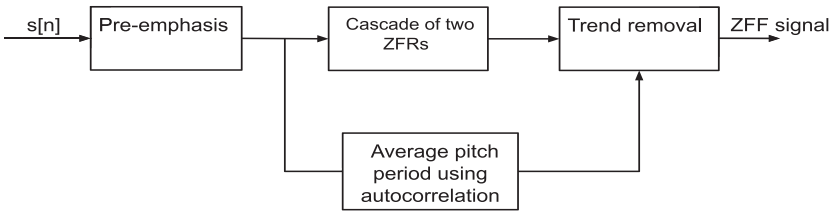

Fig. 2. Block diagram of the ZFF method.

where $a_1 = +4, a_2 = -6, a_3 = +4, a_4 = -1$. The resulting signal $y_o[n]$ is equivalent to integrating or cumulatively summing (in the discrete-time domain) the signal four times, which makes the signal grow as a polynomial function of time. The growing trend is removed from $y_o[n]$ by subtracting the local mean computed over the average pitch period. The trend removed signal ($y[n]$) is referred to as the zero frequency filtered signal (ZFFS) and is given by:

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

Here $2N + 1$ corresponds to the number of samples used to remove the trend. The ZFF signal can be regarded as an approximate glottal source waveform in analysing glottal source characteristics (Kadiri and Alku, 2019a; 2019c; Murty and Yegnanarayana, 2008). The positive-to-negative zero-crossings (PNZCs) correspond to GCIs by considering the negative polarity of the signal (Murty and Yegnanarayana, 2008; Kadiri and Yegnanarayana, 2017). The steps involved in the ZFF method are shown in Fig. 2.

To illustrate examples of glottal source waveforms computed by QCP and ZFF, a segment of speech signal (Fig. 3) and the corresponding simultaneously recorded NSA signal (Fig. 4) are considered. In both figures, the input (speech vs. NSA signal) is shown in (a), and the glottal source waveforms computed by QCP and ZFF are shown in (b) and (c), respectively.

3. Extraction of glottal features

This section describes the extraction of features from the glottal source waveforms computed using the QCP and ZFF methods. In addition, the extraction of MFCCs from the glottal source waveforms is described.

3.1. Glottal features derived using the QCP method

Different methods have been developed for the parameterization of the glottal source waveform and they can be grouped into two categories: time-domain glottal features and frequency-domain glottal features.

3.1.1. Time-domain glottal features

Time-domain glottal source waveforms can be parameterized using amplitude-based and time-based features (Airas, 2008; Alku, 2011). The amplitude quotient (AQ) (Alku and Vilkman, 1996; Alku et al., 2006) and the normalized amplitude quotient

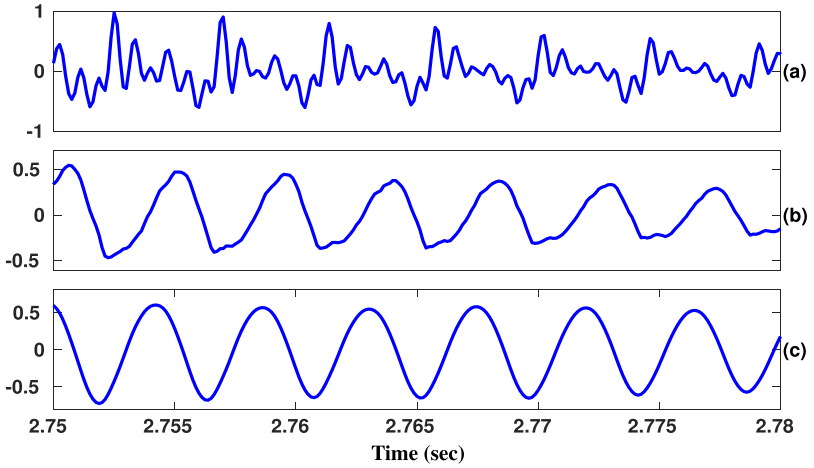

Fig. 3. Glottal source waveforms derived using the QCP and ZFF methods: (a) speech signal, (b) glottal source waveform estimated by QCP, and (c) approximate glottal source waveform estimated by ZFF (reversed in polarity for visualization purpose). The y-axes are normalized using the minimum and maximum values of the signals and expressed in arbitrary units.

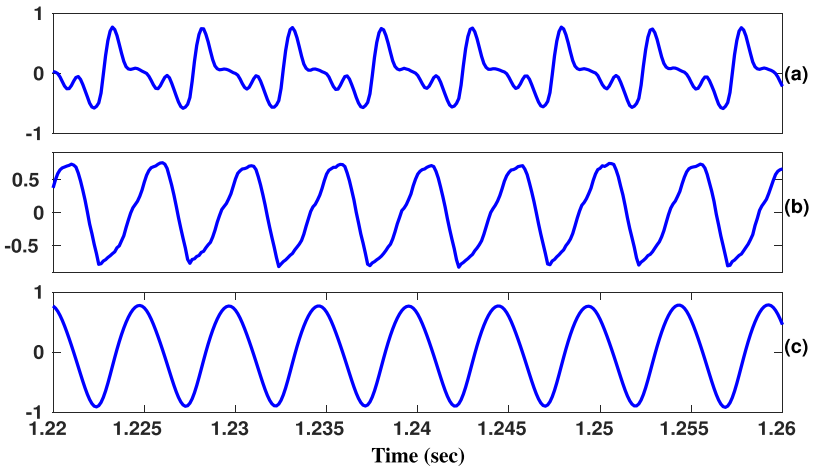

Fig. 4. Glottal source waveforms derived using the QCP and ZFF methods: (a) NSA signal, (b) glottal source waveform estimated by QCP, and (c) approximate glottal source waveform estimated by ZFF (reversed in polarity for visualization purpose). The y-axes are normalized using the minimum and maximum values of the signals and expressed in arbitrary units.

(NAQ) (Alku et al., 2002a)) are the most widely used amplitude-based glottal features utilising amplitude values of the glottal flow and its derivative in the parameterization (Fant, 1995; Alku and Vilkman, 1996; Alku et al., 2002a). NAQ was shown to be strongly correlated with the closing quotient, which has been used widely in the study of voice quality (Alku et al., 2002a). In time-based features, the classical approach is to compute time-duration ratios between the various phases (opening phase, closing phase, and closed phase) of the glottal flow pulse. These measures use the critical time instants, such as the GCI, primary and secondary glottal opening, the instant of minimum and maximum glottal flow from the glottal source waveform. Detecting the critical time instants is often difficult and to overcome this problem time-based features are sometimes computed by replacing the true closure and opening instants by the time instants when the glottal source crosses a certain level. This level is set to a value based on the maximum and minimum amplitude of the glottal pulse during the fundamental period (Alku, 2011).

3.1.2. Frequency-domain glottal features

Frequency-domain features are computed from the spectrum of the glottal source waveform to measure the slope of the spectrum. Several studies have quantified the spectral slope by using the amplitude of F0 and its harmonics. Features such as the amplitude difference between F0 and the next harmonic (H1-H2) (Titze and Sundberg, 1992), the harmonic richness factor (HRF) (Childers and Lee, 1991), and the parabolic spectral parameter (PSP) (Alku et al., 1997) are most widely used. HRF is computed as the ratio of the sum of the amplitudes of the harmonics above F0 and the amplitude of F0. PSP is derived by fitting a parabola to low frequencies of the glottal source spectrum (Alku et al., 1997).

In total, 12 glottal features (9 time-domain features and 3 frequency-domain features, listed in Table 1) are derived in this study to characterize the glottal source waveforms estimated by the QCP method (Airas, 2008). The glottal features are extracted using the APARAT Toolbox (Airas, 2008).

Table 1
Time-domain and frequency-domain glottal features derived from glottal source waveforms estimated by the QCP method.

	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

3.2. Glottal features derived using the ZFF method

To quantify the glottal source characteristics from the ZFF signal, the following four features are computed in the present study: *the slope of the ZFFS* (ZFFS slope), *the energy of the ZFFS* (ZFFS energy), the *energy of excitation* (EoE), and the *loudness measure* (Loudness). These features have been shown to be useful for discriminating phonation types, emotions and voice pathologies (Kadiri and Yegnanarayana, 2018a; Kadiri et al., 2015). By denoting GCIs as $\mathcal{G} = \{g_1, g_2, \ldots, g_M\}$ (M is the number of GCIs), these four features are computed as follows.

ZFFS slope is the slope of the ZFFS around the c^{th} GCI and is given by:

$$ZFFSslope_{g_c} = \Big|y[g_c + 1] - y[g_c - 1]\Big|, c = 1, 2, ..., M. \tag{3}$$

This feature was shown to be useful in the analysis and classification of phonation type in speech, singing and emotions in Gangamohan et al. (2013) and Kadiri et al. (2015, 2020). Similarly to NAQ, the ZFFS slope shows a decreasing trend when the phonation type changes from breathy to modal and then to pressed due to the increasing trend of the closed phase (Airas and Alku, 2007; Alku et al., 2002a; Kadiri et al., 2020). In analysing vocal emotions, the ZFFS slope was shown to be large for low arousal emotions and small for high arousal emotions (Gangamohan et al., 2013; Kadiri et al., 2015).

ZFFS energy is the energy of $y[n]$ over a window of L samples around the c^{th} GCI and is given by:

$$ZFFS\ energy_{g_c} = \frac{1}{L}\sum_{i=-L/2}^{L/2} y^2[g_c + i], c = 1, 2, ..., M. \tag{4}$$

As the ZFF signal is a low-pass filtered signal, the value of the ZFFS energy reflects the amount of low-frequency information in the signal. The ZFFS energy shows a decreasing trend when the phonation type changes from breathy to pressed, which depicts the low-frequency contents of the glottal source spectrum when the phonation type changes from breathy to pressed (Kadiri et al., 2020).

EoE is derived from the Hilbert envelope ($he[n]$) of the LP residual of the input signal over a 1 ms region around the c^{th} GCI (Kadiri et al., 2015) and is computed as follows:

$$EoE_{g_c} = \frac{1}{2K+1}\sum_{i=-K}^{K} he^2[g_c + i], c = 1, 2, ..., M, \tag{5}$$

where 2K+1 corresponds to the samples in the 1 ms window. This feature was shown to capture the changes in vocal effort (Gangamohan et al., 2013; Kadiri et al., 2015), where EoE was generally large for high arousal emotions and small for low arousal emotions. EoE shows an increasing trend when the phonation type changes from breathy to pressed, indicating increased vocal effort (Kadiri et al., 2020).

Loudness is the ratio between the standard deviation (σ_{g_c}) and mean (μ_{g_c}) of the samples of $he[n]$ in a 1 ms window around the c^{th} GCI and is given by:

$$Loudness_{g_c} = \frac{\sigma_{g_c}}{\mu_{g_c}}, c = 1, 2, ..., M. \tag{6}$$

This measure was shown to indicate the abruptness of glottal closure (Seshadri and Yegnanarayana, 2009). Loudness shows an increasing trend when the phonation type changes from breathy to pressed, indicating the increase in sharpness of glottal closure (Kadiri et al., 2020).

The steps involved in the deriving the glottal source features from the ZFF method are shown in Fig. 5.

3.3. Extraction of MFCCs from glottal waveforms

Experiments reported in Kadiri and Alku (2019c,a) have shown that the features derived from the glottal source spectrum have better discrimination capability compared to the time-domain features. For an illustration, Figs. 6 and 7 show spectrograms of glottal source waveforms estimated using the QCP and ZFF methods, respectively, from NSA signals (vowel [a]) in three phonation types (breathy, modal and pressed). From the figures, it can be seen that there are clear variations in the harmonic structure between the three phonation types. Pressed phonation exhibits a richer harmonic content compared to breathy and modal voices. This observation also holds for glottal source waveforms estimated from speech signals using the QCP and ZFF methods. In order to capture these spectral variations in a compact form, MFCCs are derived from the spectra of the glottal source waveforms. It is to be noted that the extraction of MFCCs used here is equivalent to the widely used conventional MFCC feature extraction approach (Davis and Mermelstein, 1980), except that the glottal source estimated from the speech/NSA signal is used as the input to the MFCC chain instead of the speech signal. Fig. 8 shows the steps involved in the extraction of MFCCs from the glottal source waveforms obtained by the QCP and ZFF methods, and the corresponding features are referred to as QCP-MFCC and ZFF-MFCC, respectively.

4. Experimental protocol

This section describes the database, the feature sets designed for the classification experiments and the details of the classifier.

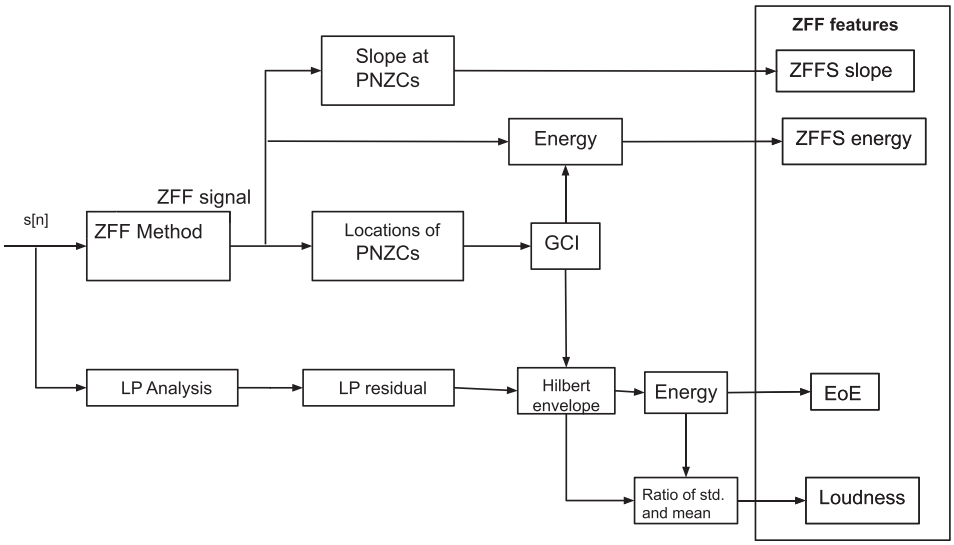

Fig. 5. Block diagram describing the steps involved in the extraction of glottal features using the ZFF method..

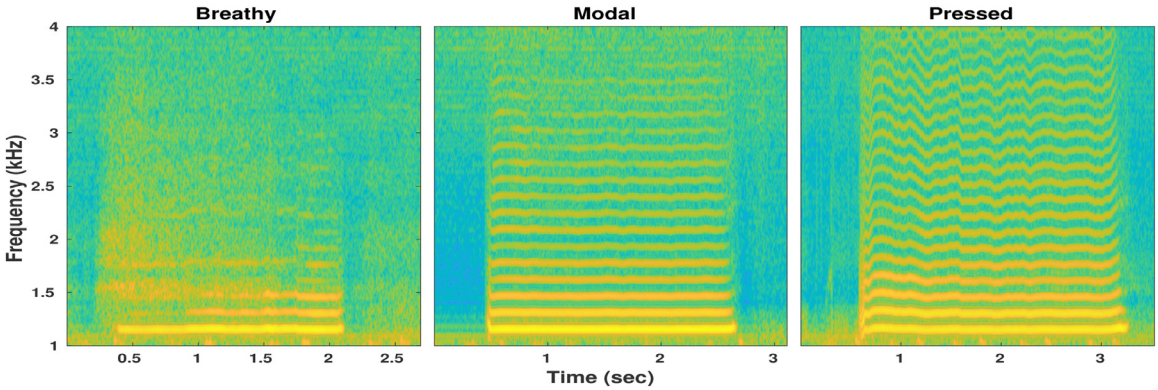

Fig. 6. Spectrograms of glottal source waveforms estimated from the NSA signal using the QCP method for breathy, modal and pressed [a] vowels.

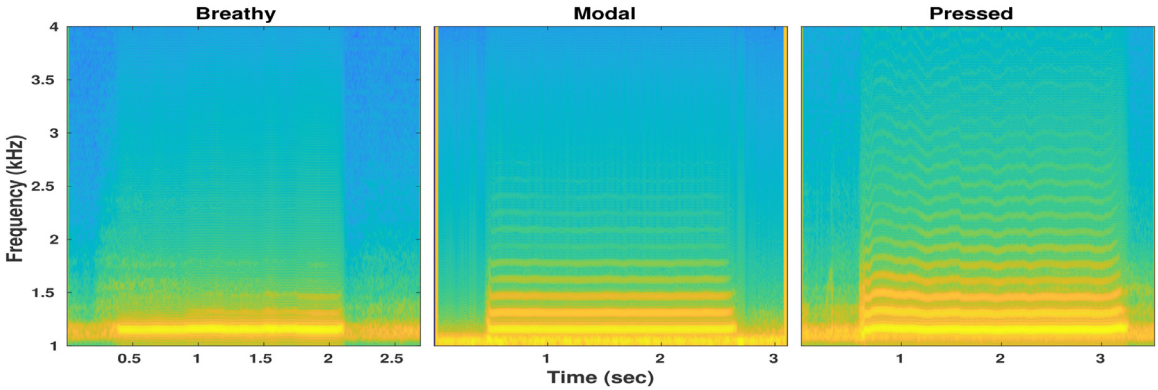

Fig. 7. Spectrograms of glottal source waveforms estimated from the NSA signal using the ZFF method for breathy, modal and pressed [a] vowels.

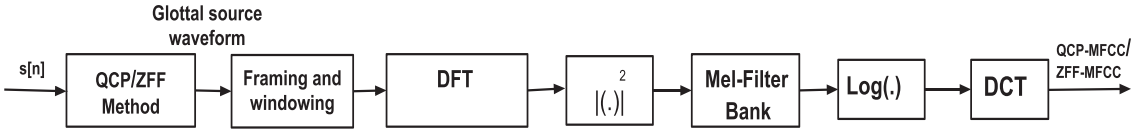

Fig. 8. Extraction of MFCCs from the glottal source waveforms computed by the QCP and ZFF methods.

4.1. Database

The database used in the present study consists of five vowels ([*a*] in the word "father", [æ] in the word "cat", [*e*] in the word "bed", [*i*] in the word "heat" and [*u*] in the word "food") uttered in three phonation types (breathy, modal and pressed) by 31 native Canadian English female speakers, aged between 18 and 40 years (Lei et al., 2019). The database consists of simultaneous recordings of the acoustic speech signal (captured by microphone) and the NSA signal. The protocol for each participant began with a training session, where the participants were instructed by an speech language pathologist (SLP) to practice the production of the three phonation types. During the recording session, the corresponding utterance was repeated until the target phonation type (judged by the SLP) was achieved. The entire recording session took approximately 30 min for each speaker. Each vowel was uttered three times using the three phonation types, resulting in a total of $5 \cdot 3 \cdot 3 \cdot 31 = 1395$ vowels. The database was originally recorded using a sampling frequency of 44.1 kHz but the data was down-sampled to 16 kHz for the purposes of this study. All the recorded speech signals were perceptually assessed independently by five SLPs, and the obtained scores were analysed in terms of their inter-rater and intra-rater reliability. This process resulted in 952 samples (out of 1395 samples) which were considered to represent reliably the corresponding phonation type. From these 952 samples, 395 are breathy, 285 are modal and 272 are pressed. The entire duration of the data is around 52 min. More details of the database can be found in Lei et al. (2019).

It is worth noting that the database used in the current study is much smaller than databases currently used in speech technology areas such as speech recognition, speaker recognition, and speech synthesis. However, to the best of our knowledge, the selected database is the only phonation type database, which includes simultaneous recordings of speech and NSA signals and which is currently available for research purposes.

4.2. Feature sets

In total, four glottal feature sets and one MFCC feature set were designed for the phonation type classification in the present study. All these sets were computed to express the two input signals that were of interest in the study, the speech signal and the NSA signal. The *first* feature set consists of the following 12 glottal features derived using the QCP method: OQ1, OQ2, NAQ, ClQ, SQ1, SQ2, AQ, QOQ, OQa (time-domain features) and H1-H2, PSP, HRF (frequency-domain features). This set is referred to as QCP-1D (with reference to the use of QCP and 1-dimensional features). All these features were extracted for every glottal cycle with QCP using Hamming-windowed 25 ms frames with a 5 ms shift and a vocal tract order of 30. The *second* set consists of the following 4 glottal features derived using the ZFF method: ZFFS slope, EoE, Loudness and ZFFS energy. This set is referred to as ZFF-1D. All these features were computed around GCIs. EoE and loudness measure were computed from a 1 ms region of the Hilbert envelope of the LP residual (computed using an order of 12) around each GCI. The *third* set consists of the MFCC features derived from the glottal source waveforms estimated by QCP. This set is referred to as the QCP-MFCC features. The *fourth* set consists of the MFCC features derived from the glottal source waveforms estimated by ZFF. This set is referred to as the ZFF-MFCC features. In addition to the glottal sets above, we computed as the *fifth* set MFCCs directly from the input signal (i.e. the speech signal and the NSA signal). All the MFFC-based feature sets were computed using 25 ms Hamming-windowed frames with a 5 ms frame shift and using 13 static coefficients and their delta & double-delta coefficients yielding 39-dimensional feature vectors. The number of mel-filter banks used was 40 and the DFT size was 1024.

Experiments were carried out with the individual feature sets as well as with combinations of the feature sets to analyze complementary information between the feature sets. In total, nine feature sets were investigated, out of which five were individual feature sets (denoted by FS-1 to FS-5) and four were combinations of feature sets (denoted by FS-6 to FS-9). In the combination of the feature sets, complementary information was studied both between the glottal feature sets (FS-6 to FS-8) and between the glottal and MFCC feature sets (FS-9). The last combined set (FS-9) was built by combining the conventional MFCC features with the proposed glottal source feature set that yielded the highest accuracy. In other words, FS-9 included FS-5 combined with the best set of glottal source features (from FS-1 to FS-4 and from FS-6 to FS-8). In summary, the 9 feature sets used in the current study are listed below.

- FS-1: QCP-1D
- FS-2: ZFF-1D
- FS-3: QCP-MFCC
- FS-4: ZFF-MFCC
- FS-5: MFCCs
- FS-6: Combination of the QCP-based sets (QCP-1D, QCP-MFCC)
- FS-7: Combination of the ZFF-based sets (ZFF-1D, ZFF-MFCC)
- FS-8: Combination of all glottal feature sets (QCP-1D, QCP-MFCC, ZFF-1D, ZFF-MFCC)
- FS-9: Combination of the best of (FS-1, FS-2, FS-3, FS-4, FS-6, FS-7, FS-8) and FS-5.

4.3. Classifier

Support vector machine (SVM) with radial basis function kernel is used as a classifier. The Scikit-learn Python library (Pedregosa et al., 2011; Chang and Lin, 2011) was used to implement the SVM classifier. The default values of the SVM classifier are

used for all hyper-parameters. It is known that the SVM classifier is effective particularly in cases where a small amount of training data is available as in the present study (Kane and Gobl, 2013; Borsky et al., 2017b). Experiments are conducted using leave one speaker out (LOSO) strategy. That is, one speaker data out of 31 speakers (consisting of all three phonation types) is used for testing, and remaining 30 speakers data (consisting of all three phonation types) is used for training. Classification accuracies for each of the speaker is first computed, and finally the mean and standard deviation of these accuracies are computed.

5. Results

This section reports the study results by first describing the classification accuracies obtained using the designed features sets and then reporting confusion matrices for the two input signals (speech vs. NSA).

Results of the phonation type classification experiments are shown in terms of the mean and standard deviation of the classification accuracy in Table 2. From the table, it can be clearly seen as the general trend that the classification accuracy is higher when the NSA signal is used instead of the speech signal as input: the accuracy is higher in NSA compared to speech in all nine feature sets. This observation is as expected in the study hypothesis. This result suggests that in comparison to the acoustic speech signal, the NSA signal includes more information about the functioning of the vocal folds when speakers change their phonation type from breathy to modal and then to pressed. In the glottal feature sets (FS-1 to FS-4), the ZFF-MFCC features (FS-4) show the best performance for the NSA signal and the QCP-1D (FS-1) features show the best performance for the speech signal. In comparing the QCP-based feature sets (FS-1 and FS-3) for speech, the QCP-1D features (FS-1) show better performance than the QCP-MFCC features (FS-3). In the ZFF-based features (FS-2 and FS-4), the ZFF-MFCC features (FS-4) are better than the ZFF-1D (FS-2) features both for the speech signal and for the NSA signal. It is interesting to observe that the conventional MFCC features (FS-5) perform better than any of the glottal features both in the speech signal and in the NSA signal. By comparing FS-6 to FS-1 and FS-3, it can be seen that the combination of the QCP-based feature sets improved the accuracy for the NSA signal and nearly the same happened for the speech signal. In addition, accuracy also improved when the ZFF-based feature sets were combined as can be seen by comparing FS-7 to FS-2 and FS-4. When both the QCP-based and ZFF-based feature sets were combined (FS-8), accuracy improved further both in the speech signal and in the NSA signal. It is interesting to note that the combined QCP-based and ZFF-based feature sets (FS-8) perform better than conventional MFCC features (FS-5) both in the speech signal and in the NSA signal. Finally, the combination of the conventional MFCCs with all the glottal feature sets (FS-9) lead to a further improvement in accuracy both in speech and NSA indicating the existence of complementary information between the conventional MFCC features and glottal features.

Class-wise accuracy was analysed in terms of confusion matrices using the combined feature sets (i.e. the sets from FS-6 to FS-9). Table 3 shows the confusion matrices for the speech and NSA signal for the FS-6 to FS-9 feature sets. In the case of the speech signal, it can be clearly seen for all the feature sets that there exists confusion between breathy and modal voices as well as between modal and pressed voices. The same is true also for the NSA signal, even though there is an improvement in accuracy in all phonation types. It can also be observed that the class-wise accuracies show an increasing trend for the feature sets from FS-6 to FS-9, especially for modal voice, and also to some extent for breathy and pressed voices. It should be noted that even though there is an improvement in classification accuracy from FS-6 to FS-9, modal voices are confused with breathy and pressed voices which makes the overall accuracy lower both in the speech and NSA signal. These observations indicate that there is a need for further investigations to develop features that reflect differences in voice production characteristics between the three phonation types. It addition, further research is needed to better understand the complexity of the classification problem by evaluating overlapping between the classes. In order to study the complexity of the phonation type classification task, the techniques described in Lorena et al. (2019), for example, could be used.

Table 2
Phonation type classification accuracy (mean and standard deviation) for the speech signal and the NSA signal for individual feature sets and combinations of feature sets.

Feature set	**Speech** [%]	**NSA** [%]
FS-1	70.9 ± 3.4	74.5 ± 3.9
FS-2	66.3 ± 2.1	73.3 ± 0.9
FS-3	63.4 ± 2.0	71.6 ± 5.7
FS-4	67.4 ± 3.9	76.7 ± 2.5
FS-5	75.8 ± 1.3	84.0 ± 3.4
FS-6	70.2 ± 3.6	78.5 ± 4.4
FS-7	73.0 ± 4.2	80.4 ± 2.4
FS-8	76.9 ± 4.6	84.9 ± 2.2
FS-9	80.6 ± 2.2	86.9 ± 2.7

Table 3
Confusion matrices in phonation type classification from speech and NSA signals using the QCP-based feature set (FS-6), the ZFF-based feature set (FS-7), the combination of the QCP-based and ZFF-based feature sets (FS-8), and the combination of conventional MFCCs and all glottal feature sets (FS-9). Here B, M and P refer to breathy, modal and pressed voices, respectively.

Feature set	**Speech** [%]				**NSA** [%]			
FS-6		B	M	P		B	M	P
	B	**82.3**	11.9	5.8	B	**87.8**	8.6	3.4
	M	31.6	**48.4**	20.0	M	23.5	**61.8**	14.7
	P	13.6	11	**75.4**	P	8.1	9.5	**82.4**
FS-7		B	M	P		B	M	P
	B	**85.1**	10.1	4.8	B	**85.8**	8.9	5.3
	M	30.9	**50.9**	18.2	M	19.6	**65.6**	14.8
	P	8.4	12.9	**78.7**	P	4.0	8.1	**87.9**
FS-8		B	M	P		B	M	P
	B	**87.8**	9.6	2.6	B	**90.6**	6.3	3.1
	M	28.1	**55.8**	16.1	M	14.8	**71.2**	14.0
	P	7.7	9.2	**83.1**	P	2.2	6.6	**91.2**
FS-9		B	M	P		B	M	P
	B	**88.4**	8.6	3.0	B	**92.4**	5.3	2.3
	M	22.5	**63.2**	14.3	M	12.3	**75.4**	12.3
	P	6.6	5.9	**87.5**	P	4.0	4.8	**91.2**

6. Conclusions

In this article, classification of phonation type into three classes (breathy, modal, pressed) was studied using the acoustic speech signal and the simultaneously recorded NSA signal. Features describing the glottal source were derived using two signal processing methods, QCP and ZFF. QCP estimates glottal source waveforms based on the source-filter decomposition while ZFF computes source waveforms without explicitly using the source-filter decomposition. Using the glottal source waveforms obtained by these two methods, several scalar glottal features were computed. In addition, the glottal source waveforms were parameterized using MFCCs. Classification experiments using different glottal features with SVM revealed that the NSA signal has a better capability to discriminate the three phonation types compared to using the acoustic speech signal. It was also observed that there exists complementary information between the glottal features computed by QCP and ZFF. Furthermore, it was observed that the classification accuracy improved both for the speech signal and the NSA signal when the glottal features were combined with the conventional MFCCs, indicating complementary information between these feature sets. Finally, we would like to point out that the classification experiments conducted in the current study were all based on treating the two signals separately, that is, we did not merge information extracted from the acoustical speech signal to information extracted from the NSA signal. Since both the speech signal and the NSA signal carry valuable information related to phonation type, a potential topic for future studies is to investigate how merging information from these two signals could further improve the classification performance.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This study was funded by the Academy of Finland (Project No. 312490 and 330139).

References

Afshan, A., Guo, J., Park, S.J., Ravi, V., Flint, J., Alwan, A., 2018. Effectiveness of voice quality features in detecting depression. In: Proc. INTERSPEECH, pp. 1676–1680.
Airaksinen, M., Raitio, T., Story, B., Alku, P., 2014. Quasi closed phase glottal inverse filtering analysis with weighted linear prediction. IEEE/ACM Trans. on Audio, Speech, and Lang. Process. 22 (3), 596–607.
Airas, M., 2008. Tkk aparat: an environment for voice inverse filtering and parameterization. Logopedics Phoniatrics Vocol. 33 (1), 49–64.
Airas, M., Alku, P., 2007. Comparison of multiple voice source parameters in different phonation types. In: Proc. INTERSPEECH, pp. 1410–1413.
Alku, P., 2011. Glottal inverse filtering analysis of human voice production-a review of estimation and parameterization methods of the glottal excitation and their applications. Sadhana 36 (5), 623–650.
Alku, P., Airas, M., Björkner, E., Sundberg, J., 2006. An amplitude quotient based method to analyze changes in the shape of the glottal pulse in the regulation of vocal intensity. J. Acoust. Soc. Am. 120, 1052–1062.
Alku, P., Bäckström, T., Vilkman, E., 2002. Normalized amplitude quotient for parameterization of the glottal flow. J. Acoust. Soc. Am. 112, 701–710.
Alku, P., Pohjalainen, J., Vainio, M., Laukkanen, A.-M., Story, B.H., 2013. Formant frequency estimation of high-pitched vowels using weighted linear prediction. J. Acoust. Soc. Am. 134 (2), 1295–1313.
Alku, P., Strik, H., Vilkman, E., 1997. Parabolic spectral parameter - a new method for quantification of the glottal flow. Speech Commun. 22 (1), 67–79.

Alku, P., Vilkman, E., 1996. Amplitude domain quotient for characterization of the glottal volume velocity waveform estimated by inverse filtering. Speech Commun. 18, 131–138.
Alku, P., Vintturi, J., Vilkman, E., 2002. Measuring the effect of fundamental frequency raising as a strategy for increasing vocal intensity in soft, normal and loud phonation. Speech Commun. 38 (3–4), 321–334.
Birkholz, P., Martin, L., Willmes, K., Kröger, B.J., Neuschaefer-Rube, C., 2015. The contribution of phonation type to the perception of vocal emotions in german: an articulatory synthesis study. J. Acoust. Soc. Am. 137 (3), 1503–1512.
Borsky, M., Cocude, M., Mehta, D.D., Zañartu, M., Gudnason, J., 2017. Classification of voice modes using neck-surface accelerometer data. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 5060–5064.
Borsky, M., Mehta, D.D., Van Stan, J.H., Gudnason, J., 2017. Modal and nonmodal voice quality classification using acoustic and electroglottographic features. IEEE/ACM Trans. Audio, Speech, and Lang. Process. 25 (12), 2281–2291.
Campbell, N., Mokhtari, P., 2003. Voice quality: the 4th prosodic dimension. In: Proc. ICPhS, pp. 2417–2420.
Chang, C.-C., Lin, C.-J., 2011. LIBSVM: A library for support vector machines. ACM Trans. Intel. Syst.Technol. (TIST) 2 (3), 1–27.
Childers, D.G., Lee, C.K., 1991. Vocal quality factors: analysis, synthesis, and perception. J. Acoust. Soc. Am. 90 (5), 2394–2410.
Coleman, R.F., 1988. Comparison of microphone and neck-mounted accelerometer monitoring of the performing voice. J. Voice 2 (3), 200–205.
Cortés, J.P., Espinoza, V.M., Ghassemi, M., Mehta, D.D., Van Stan, J.H., Hillman, R.E., Guttag, J.V., Zañartu, M., 2018. Ambulatory assessment of phonotraumatic vocal hyperfunction using glottal airflow measures estimated from neck-surface acceleration. PLoS ONE 13 (12), 1–22.
Davis, S., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. 28 (4), 357–366. https://doi.org/10.1109/TASSP.1980.1163420.
ud Dowla Khan, S., 2012. The phonetics of contrastive phonation in Gujarati. J. Phon. 40 (6), 780–795.
Drugman, T., Alku, P., Alwan, A., Yegnanarayana, B., 2014. Glottal source processing: from analysis to applications. Comput. Speech Lang. 28 (5), 1117–1138.
Esposito, C.M., 2010. The effects of linguistic experience on the perception of phonation. J. Phon. 38 (2), 306–316.
Fant, G., 1995. The lf-model revisited. transformations and frequency domain analysis. Speech Transmiss. Lab. Q. Progress Status Rep. 36, 119–156.
Gangamohan, P., Kadiri, S.R., Yegnanarayana, B., 2013. Analysis of emotional speech at subsegmental level. In: Proc. INTERSPEECH, pp. 1916–1920.
Garellek, M., Samlan, R., Gerratt, B.R., Kreiman, J., 2016. Modeling the voice source in terms of spectral slopes. J. Acoust. Soc. Am. 139 (3), 1404–1410.
Ghassemi, M., Van Stan, J.H., Mehta, D.D., Zañartu, M., Cheyne II, H.A., Hillman, R.E., Guttag, J.V., 2014. Learning to detect vocal hyperfunction from ambulatory neck-surface acceleration features: initial results for vocal fold nodules. IEEE Trans. Biomed. Eng. 61 (6), 1668–1675.
Gobl, C., Ní Chasaide, A., 2003. The role of voice quality in communicating emotion, mood and attitude. Speech Commun. 40 (1–2), 189–212.
Gordon, M., Ladefoged, P., 2001. Phonation types: a cross-linguistic overview. J. Phon. 29 (4), 383–406.
Gowda, D., Kurimo, M., 2013. Analysis of breathy, modal and pressed phonation based on low frequency spectral density. In: Proc. INTERSPEECH, pp. 3206–3210.
Grichkovtsova, I., Morel, M., Lacheret, A., 2012. The role of voice quality and prosodic contour in affective speech perception. Speech Commun. 54 (3), 414–429.
Hillenbrand, J., Cleveland, R.A., Erickson, R.L., 1994. Acoustic correlates of breathy vocal quality. J. Speech Lang. Hear. Res. 37 (4), 769–778.
Ito, M., 2004. Politeness and voice quality-the alternative method to measure aspiration noise. In: Proc. Speech Prosody.
Kadiri, S.R., Alku, P., 2019. Analysis and detection of pathological voice using glottal source features. IEEE J. Sel. Top. Signal Process. 14 (2), 367–379.
Kadiri, S.R., Alku, P., 2019. Mel-frequency cepstral coefficients derived using the zero-time windowing spectrum for classification of phonation types in singing. J. Acoust. Soc. Am. 146 (5), EL418–EL423.
Kadiri, S.R., Alku, P., 2019. Mel-frequency cepstral coefficients of voice source waveforms for classification of phonation types in speech. Proc. Interspeech 2508–2512.
Kadiri, S.R., Alku, P., Yegnanarayana, B., 2020. Analysis and classification of phonation types in speech and singing voice. Speech Commun. 118, 33–47.
Kadiri, S.R., Gangamohan, P., Gangashetty, S.V., Yegnanarayana, B., 2015. Analysis of excitation source features of speech for emotion recognition. In: Proc. INTERSPEECH, pp. 1324–1328.
Kadiri, S.R., Yegnanarayana, B., 2017. Speech polarity detection using strength of impulse-like excitation extracted from speech epochs. ICASSP, pp. 5610–5614.
Kadiri, S.R., Yegnanarayana, B., 2018. Analysis and detection of phonation modes in singing voice using excitation source features and single frequency filtering cepstral coefficients (SFFCC). In: Proc. INTERSPEECH, pp. 441–445.
Kadiri, S.R., Yegnanarayana, B., 2018. Breathy to tense voice discrimination using zero-time windowing cepstral coefficients (ztwccs). In: Proc. INTERSPEECH, pp. 232–236.
Kane, J., Gobl, C., 2013. Wavelet maxima dispersion for breathy to tense voice discrimination. IEEE Trans. Audio, Speech & Lang. Process. 21 (6), 1170–1179.
Klatt, D.H., Klatt, L.C., 1990. Analysis, synthesis, and perception of voice quality variations among female and male talkers. J. Acoust. Soc. Am. 87 (2), 820–857.
Kreiman, J., Park, S.J., Keating, P.A., Alwan, A., 2015. The relationship between acoustic and perceived intraspeaker variability in voice quality. In: Proc. INTERSPEECH, pp. 2357–2360.
Kreiman, J., Shue, Y.-L., Chen, G., Iseli, M., Gerratt, B.R., Neubauer, J., Alwan, A., 2012. Variability in the relationships among voice quality, harmonic amplitudes, open quotient, and glottal area waveform shape in sustained phonation. J. Acoust. Soc. Am. 132, 2625–2632.
Kuang, J., Keating, P., 2014. Vocal fold vibratory patterns in tense versus lax phonation contrasts. J. Acoust. Soc. Am. 136 (5), 2784–2797.
Ladefoged, P., Maddieson, I., Jackson, M., 1988. Investigating phonation types in different languages. Vocal Physiology: Voice Production, Mechanisms and Functions. New York: Raven Press.
Laver, J., 1980. The Phonetic Description of Voice Quality. Cambridge University Press, Cambridge.
Lei, Z., Kennedy, E., Fasanella, L., Li-Jessen, N.Y.-K., Mongeau, L., 2019. Discrimination between modal, breathy and pressed voice for single vowels using neck-surface vibration signals. Appl. Sci. 9 (7), 1505.
Lin, J.Z., Espinoza, V.M., Marks, K.L., Zañartu, M., Mehta, D.D., 2020. Improved subglottal pressure estimation from neck-surface vibration in healthy speakers producing non-modal phonation. IEEE J. Sel. Top. Signal Process. 14 (2), 449–460.
Lorena, A.C., Garcia, L.P., Lehmann, J., Souto, M.C., Ho, T.K., 2019. How complex is your classification problem? a survey on measuring classification complexity. ACM Comput. Surv. (CSUR) 52 (5), 1–34.
Lulich, S.M., Morton, J.R., Arsikere, H., Sommers, M.S., Leung, G.K., Alwan, A., 2012. Subglottal resonances of adult male and female native speakers of american english. J. Acoust. Soc. Am. 132 (4), 2592–2602.
Mehta, D.D., Espinoza, V.M., Van Stan, J.H., Zañartu, M., Hillman, R.E., 2019. The difference between first and second harmonic amplitudes correlates between glottal airflow and neck-surface accelerometer signals during phonation. J. Acoust. Soc. Am. 145 (5), EL386–EL392.
Mehta, D.D., Van Stan, J.H., Hillman, R.E., 2016. Relationships between vocal function measures derived from an acoustic microphone and a subglottal neck-surface accelerometer. IEEE/ACM Trans. Audio Speech Lang. Process. 24 (4), 659–668.
Mehta, D.D., Van Stan, J.H., Zañartu, M., Ghassemi, M., Guttag, J.V., Espinoza, V.M., Cortés, J.P., Cheyne, H.A., Hillman, R.E., 2015. Using ambulatory voice monitoring to investigate common voice disorders: research update. Front. Bioeng. Biotechnol. 3, 155.
Murty, K.S.R., Yegnanarayana, B., 2008. Epoch extraction from speech signals. IEEE Trans. Audio Speech Lang. Process. 16 (8), 1602–1613.
Narendra, N., Alku, P., 2020. Glottal source information for pathological voice detection. IEEE Access 8, 67745–67755.
Park, S.J., Afshan, A., Chua, Z.M., Alwan, A., 2018. Using voice quality supervectors for affect identification. In: Proc. INTERSPEECH, pp. 157–161.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Édouard Duchesnay, 2011. Scikit-learn: machine learning in python. J. Mach. Learn. Res. 12 (85), 2825–2830.
Pietrowicz, M., Hasegawa-Johnson, M., Karahalios, K.G., 2017. Acoustic correlates for perceived effort levels in male and female acted voices. J. Acoust. Soc. Am. 142 (2), 792–811.
Rendon, D.B., Ojeda, J.L.R., Foix, L.F.C., Morillo, D.S., Fernández, M.A., 2007. Mapping the human body for vibrations using an accelerometer. 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE, pp. 1671–1674.
Seshadri, G., Yegnanarayana, B., 2009. Perceived loudness of speech based on the characteristics of glottal excitation source. J. Acoust. Soc. Am. 126, 2061–2071.

Stevens, K.N., Kalikow, D.N., Willemain, T.R., 1975. A miniature accelerometer for detecting glottal waveforms and nasalization. J. Speech Hear. Res. 18 (3), 594–599.
Sun, X., Jiang, Y., Li, W., 2020. Residual attention based network for automatic classification of phonation modes. 2020 IEEE International Conference on Multimedia and Expo (ICME). IEEE, pp. 1–6.
Swerts, M., Veldhuis, R.N.J., 2001. The effect of speech melody on voice quality. Speech Commun. 33 (4), 297–303.
Titze, I., Sundberg, J., 1992. Vocal intensity in speakers and singers. J. Acoust. Soc. Am. 91, 2936–2946.
Titze, I.R., 2000. Principles of Voice Production (second printing). Iowa City, IA: National Center for Voice and Speech.
Titze, IR, Svec, JG, Popolo, PS, 2003. Vocal dose measures: quantifying accumulated vibration exposure in vocal fold tissues. J. Speech Lang. Hear. Res. 46 (4), 919–932.
Titze, I.R., Hunter, E.J., 2015. Comparison of vocal vibration-dose measures for potential-damage risk criteria. J. Speech Lang. Hear. Res. 58 (5), 1425–1439.
Van Stan, J.H., Mehta, D.D., Zeitels, S.M., Burns, J.A., Barbu, A.M., Hillman, R.E., 2015. Average ambulatory measures of sound pressure level, fundamental frequency, and vocal dose do not differ between adult females with phonotraumatic lesions and matched control subjects. Ann. Otol. Rhinol. Laryngol. 124 (11), 864–874.
Wong, D., Markel, J., Gray, A., 1979. Least squares glottal inverse filtering from the acoustic speech waveform. IEEE Trans. Audio Speech Signal Process. 27, 350–355.
Yanushevskaya, I., Gobl, C., Chasaide, A.N., 2005. Voice quality and f0 cues for affect expression: implications for synthesis. Ninth European Conference on Speech Communication and Technology, pp. 1849–1852.
Zañartu, M., Ho, J.C., Mehta, D.D., Hillman, R.E., Wodicka, G.R., 2013. Subglottal impedance-based inverse filtering of voiced sounds using neck surface acceleration. IEEE Trans. Audio Speech Lang. Process. 21 (9), 1929–1939.