Kadiri, Sudarsana Reddy; Alku, Paavo; Yegnanarayana, Bayya

Analysis and classification of phonation types in speech and singing voice

Published in:
Speech Communication

DOI:
10.1016/j.specom.2020.02.004

Published: 01/04/2020

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Analysis and Classification of Phonation Types in Speech and Singing Voice

Sudarsana Reddy Kadiri[1]*, Paavo Alku[1], B. Yegnanarayana[2]

[1]*Department of Signal Processing and Acoustics, Aalto University, Finland.*
[2]*Speech Processing Laboratory, IIIT-Hyderabad, India.*

Abstract

Both in speech and singing, humans are capable of generating sounds of different phonation types (e.g., breathy, modal and pressed). Previous studies in the analysis and classification of phonation types have mainly used voice source features derived using glottal inverse filtering (GIF). Even though glottal source features are useful in discriminating phonation types in speech, their performance deteriorates in singing voice due to the high fundamental frequency of these sounds that reduces the accuracy of source-filter separation in GIF. In the present study, features describing the glottal source were computed using three signal processing methods that do not compute source-filter separation. These three methods are zero frequency filtering (ZFF), zero time windowing (ZTW) and single frequency filtering (SFF). From each method, a group of scalar features were extracted. In addition, cepstral coefficients were derived from the spectra computed using ZTW and SFF. Experiments were conducted with the proposed features to analyse and classify phonation types using three phonation types (breathy, modal and pressed) for speech and singing voice. Statistical pair-wise comparisons between the phonation types showed that most of the features were capable of separating the phonation types significantly for speech and singing voices. Classification with support vector machine classifiers indicated that the proposed features and their combinations showed improved accuracy compared to usually employed glottal source features and mel-frequency cepstral coefficients (MFCCs).

Key words: Phonation type, Voice quality, Singing voice, Glottal source, Glottal inverse filtering, Zero frequency filtering (ZFF), Zero time windowing (ZTW) and Single frequency filtering (SFF).

1. Introduction

Human perception of voiced sounds can be roughly described in four dimensions: pitch, loudness, vowel identity (or voiced consonant identity) and quality [1]. The last item, *quality*, is

*Corresponding author. Contact. +358 504754005.

**This work appeared in part in the Proceedings of INTERSPEECH 2018.

Email addresses: sudarsana.kadiri@aalto.fi (Sudarsana Reddy Kadiri[1]), paavo.alku@aalto.fi (Paavo Alku[1]), yegna@iiit.ac.in (B. Yegnanarayana[2])

defined in speech science as the auditory colouring of a person's voice [2]. This perceptual dimension, which is present both in voiced speech sounds and singing voices, is affected by the shape of the transglottal airflow excitation pulse–the glottal pulse–generated by the vocal folds. By regulating the activation of the laryngeal muscles and the respiratory effort, humans are capable of changing the glottal pulse shape and generating sounds of different *phonation types* such as breathy, modal and pressed [3, 4]. In this study, phonation types are analysed and automatically classified from two categories of voiced sounds–speech and singing voice. Analysis and classification of phonation types from speech can be used in different areas of speech research such as in occupational voice care [5, 6], in automatic classification of speaking styles in audiobooks [7] and in modern parametric speech synthesis [8, 9, 10]. Analysis and classification of phonation types from singing could help to diagnose voice production problems such as hypo-function and hyper-function [11, 12, 13]. In addition, since many singing students exhibit varying degrees of these malfunctions throughout the course of their studies, automatic recognition of such vocal behaviour would be useful for self-monitoring since it is difficult for students to analyze their voice production while singing.

In the following, first an overview of the previous studies on analysis and classification of phonation types in speech and singing voice is given, and then the goals of the present study are summarized. The topic is studied in the current investigation based solely on the pressure signal (either speech or singing voice) captured by a microphone.

1.1. Phonation Types in Speech

According to Ladefoged [14], phonation types occur on a continuum ranging from voiceless to glottal closure. Different phonation types such as whisper, breathy, tense, creak and falsetto can be produced by changing the activation of the laryngeal muscles [2, 8, 15]. Breathy and tense are often considered to be the two opposite ends of the voice quality continuum [3, 16]. In this study, three phonation types of speech (breathy, modal and tense) from the continuum are considered as subjects of interest. Phonation type has an important role in signalling paralinguistic information (such as mood, attitude and emotions) in speech [17, 18, 19, 20, 21]. Breathy phonation has been shown to be associated with expression of politeness, familiarity and intimacy [22]. On the other hand, tense voice has been shown to be associated with emotional states of high arousal such as anger and happiness [23, 4]. In addition to signalling paralinguistic information, phonation types are used in certain languages to generate phonological contrasts [14, 24, 25, 26, 27].

Modal phonation is typically used as the reference for comparing the produced phonation types [1, 2, 14]. In modal voice, the laryngeal tension settings are low and moderate in range [1, 2]. Vibrations of the vocal folds are mostly periodic with minimum irregularity in a sequence of glottal cycles with complete glottal closure. Breathy voice typically involves weaker levels of laryngeal tension, partial glottal closure of the glottis and often a posterior glottal chink [1, 2]. These settings lead to the generation of some amount of aspiration or turbulence noise. In addition, the harmonic structure of breathy speech is more prominent at low fundamental frequencies compared to that of modal voices [4, 28]. On the other hand, the laryngeal settings of tense voice involve an increase in the longitudinal and adductive tension [1, 2]. The sharpness of the glottal closure of tense voice results in prominent high-frequency harmonics [3, 28].

Variations in the vibration mode of the vocal folds described above result in differences in the shape of the glottal flow pulse between phonation types. The glottal flow waveform varies from a smooth, almost symmetric pulse in breathy phonation to an asymmetric pulse with sharp edges in tense phonation [16, 29]. This kind of time domain variation is reflected in the decay of the spectral envelope of the glottal source in the frequency domain [30, 31]. By taking advantage of both the time and frequency domain, many glottal source features have been developed to discriminate phonation types using flow waveforms estimated by glottal inverse filtering (GIF) [16, 3, 32]. Time domain features (such as the open quotient, the quasi-open quotient (QOQ), the closing quotient (CQ) and the speed quotient (SQ)), and amplitude-based features (such as the normalized amplitude quotient (NAQ)) have been widely used to parameterize the glottal flow and its derivative [16, 28, 33]. Examples of frequency domain features are the amplitude difference between the first and second harmonic (H1-H2) [31], the harmonic richness factor (HRF) [8] and the parabolic spectral parameter (PSP) [34], which measure the decay of the glottal flow spectrum. In [30], it was found that NAQ and H1-H2 were the best features to discriminate phonation types in speech. In addition to the parametrization methods described above, a few previous investigations (e.g., [4, 35]) have analysed phonation types by using a scheme that is based on fitting the estimated glottal flow pulse (or its derivative) with an artificial glottal source model (e.g. the Liljencrants-Fant model).

Instead of first estimating the glottal flow with GIF, some studies have measured the impact of the glottal source directly from the speech spectrum by using features such as the fundamental frequency (F0), H1-H2, H2-H4 (the amplitude difference between the second and the fourth harmonics), the spectral slope between H4 and 2 kHz, and the spectral slope between 2 kHz and 5 kHz [36, 37]. In other studies (e.g., [31, 38]), the amount of aspiration noise in speech has been analysed to detect breathy phonation based on the observation according to which the third formant region is noisier in breathy phonation compared to modal phonation.

It is known that the performance of GIF deteriorates for high-pitched speech and expressive voices [28, 33]. To overcome this, attempts have been made to directly use the time domain speech signal or the linear prediction (LP) residual to extract features describing phonation types. To capture sharp changes in the glottal closure characteristics, a feature called maximum dispersion quotient (MDQ), which uses the LP residual signal was proposed in [3]. In [30], voice source characteristics such as breathy voices showing higher open quotients and pressed voices indicating smaller open quotients, were analysed using a spectral feature called the low-frequency spectral density (LFSD). The effect of the subglottal system in the speech spectrum is larger for breathy voices owing to their higher open quotient compared to pressed voices. This results in an increase in the low-frequency spectral energy in breathy voices, typically around the region of the glottal formant. In [30], it was observed that the discrimination capabilities of LFSD and MDQ were closer to the discrimination capability of NAQ, and harmonic-to-noise ratio (HNR) seems to provide poorer discrimination between the three phonation types (breathy, modal and tense). However, HNR was shown to provide good performance in the discrimination of modal and breathy voices compared to modal and pressed voices. In [3, 32], a set of glottal source features such as NAQ, QOQ, H1-H2, PSP and MDQ, along with mel-frequency cepstral coefficients (MFCCs) derived from speech signals, were investigated for classification of phonation types in speech. The combination improves the discrimination in relation to using either glottal features

alone or MFCCs alone.

1.2. Phonation Types in Singing Voice

Voice quality, a perceptual attribute partly defined by phonation type, is one of the most salient features in singing. A singer's feelings and identity are expressed through variations in voice quality. In singing, phonation types have been categorized using four classes: breathy, modal, flow (or resonant) and pressed [39, 40, 41, 42]. The main characteristics of the four phonation types that have been studied in singing voice are described below.

Breathy phonation shows reduced vocal fold adduction and minimal vocal fold contact area, which result in laxed singing voice with a high level of turbulent noise. Therefore, HNR is generally larger in breathy singing voice compared to other phonation types [8]. In addition, it has been reported that a strong perceptual indicator of breathiness is the sensation of excessive laryngeal airflow [43]. Modal voices show full vibration of the vocal folds, along their entire length. Flow phonation is associated with a large peak-to-peak glottal flow and a small glottal leakage and is typically produced using a lowered larynx [44]. Flow phonation differs from other phonation types in the sense that its production is used as a vocal exercise, for example, in voice therapy [45, 46]. Loudness is key in using flow phonation and this phonation type enables achieving greater loudness by increasing the flow amplitude rather than by decreasing the closing phase duration as in pressed phonation [44]. Pressed phonation is associated with an elevated larynx position, which influences the vocal tract shape, and also stronger muscular tension around the vocal folds. The spectrum of a pressed singing voice shows typically a weaker fundamental and more dominating higher harmonics [47].

In [39, 47], phonation types were studied in singing using features derived from the glottal source waveform estimated with GIF. It was found that the glottal source features alone are not sufficient for classification. This is mainly due to problems of GIF for singing voice, as singing voices are typically of high pitch and source-filter coupling is strong [48, 49, 28]. In the GIF methods, the glottal source waveform is estimated by filtering the signal through the inverse of the vocal tract transfer function [28]. The glottal source estimates tend to become unreliable in the analysis of high-pitched voices (such as singing voice) due to the fact that the true glottal pulse typically shows a shorter closed phase, i.e., fewer samples for high-pitched voices and hence the estimation of the vocal tract transfer function during the closed phase is difficult for GIF [28, 48]. In other words, the coupling between the subglottal system and supraglottal system (vocal tract) is larger in the case of high-pitched voices. In [50], subglottal pressure was found to correlate with the amount of pressedness. In [51], frequency domain features–such as the spectral centroid, the spectral flux and spectral energies in different bands–were used for the classification of phonation types in singing with various glottal source features and MFCCs. In addition, features such as amplitudes of harmonics, formant frequencies and their bandwidths and amplitudes, HNR, and glottal source features were studied recently in [47]. It was observed that the largest confusions occurred between breathy and modal voices and between flow and pressed voices.

1.3. Goals of the present study

The review presented in the previous two subsections indicates that the analysis and classification of phonation types have been studied mainly by using one of the following two approaches:

(1) by first estimating the glottal flow using GIF and then parametrizing the estimated glottal flow using features such as NAQ, QOQ, H1-H2, and PSP [16, 3], or (2) by estimating the influence of glottal source on the spectrum of the speech signal directly using features such as H2-H4, spectral slope between H4 and 2 kHz, spectral slope between 2 kHz and 5 kHz, and cepstral peak prominence (CPP) [19, 20, 36, 52, 37]. Even though glottal source features have been shown to be useful for the analysis and classification of phonation types in speech (e.g., [3, 30, 53]), their performance drops when analysing phonation types in singing voice (e.g., [39, 54]). This is due to the reduced accuracy of GIF methods to conduct source-filter separation due to high pitch and strong source-filter coupling, which are typically present in singing. Furthermore, to the best of our knowledge, there are no previous studies in the analysis and automatic classification of phonation types using the same set of features for *both* speech *and* singing voice. Therefore, the first goal of the present study is to propose features that quantify the glottal excitation without conducting the source-filter separation as in GIF. Statistical distributions of these features are then analysed between different phonation types both in speech and singing voice. By taking advantage of the novel features, the second goal is to study how phonation types can be automatically classified for speech and singing voice signals.

The list of abbreviations used in this study are given in Table 1.

2. Methods of Feature Extraction

In this section, three signal processing methods that are used as the basis for feature extraction in the current study are described. These three methods are: the *zero frequency filtering* (ZFF) method [55], the *zero time windowing* (ZTW) method [56], and the *single frequency filtering* (SFF) method [57]. It is to be noted that none of these methods use the source-filter model of voice production. All three methods use the microphone pressure signal as input (i.e., either speech or singing voice), which is referred to as $s[n]$ throughout the study. This time domain input signal is normalized in amplitude to lie between -1 and +1. After the presentation of each technique, the features based on the corresponding method are described.

2.1. Zero frequency filtering (ZFF)

ZFF [55] is a straightforward signal processing method to compute an approximate voice source waveform in the time domain without explicitly using any source-filter model. ZFF is based on the observation that the impulse-like nature of the voice excitation, caused by abrupt closure of the vocal folds, is reflected across all frequencies including the zero frequency (0 Hz). In order to compute the approximate voice source signal, the differentiated microphone pressure signal ($x[n] = s[n] - s[n-1]$) is first passed through a cascade of two zero frequency resonators (a pair of poles on the unit circle at the positive real axis in the z-plane) and is given by:

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

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 microphone signal four times. Hence

Table 1: List of abbreviations.

ANOVA	Analysis of variance
BER	Band energy ratio
CoG	Center of gravity
CPP	Cepstral peak prominence
CQ	Closing quotient
DPA	Dominant peak amplitude
DPL	Dominant peak location
EoE	Energy of excitation
FS	Feature set
GCI	Glottal closure instant
GIF	Glottal inverse filtering
H	Entropy
H1-H2	Amplitude difference between the first and second harmonics
HNR	Harmonic-to-noise ratio
HRF	Harmonic richness factor
LFSD	Low-frequency spectral density
LP	Linear prediction
MDQ	Maximum dispersion quotient
MFCCs	Mel frequency cepstral coefficients
NAQ	Normalized amplitude quotient
NGD	Numerator of group delay function
PSP	Parabolic spectral parameter
QOQ	Quasi-open quotient
SF	Spectral flatness
SFF	Single frequency filtering
SFFCCs	Single frequency filtering cepstral coefficients
SQ	Speed quotient
SSG	Slope of spectral gain
SSV	Slope of spectral variance
SVM	Support vector machine
VQ	Voice quality
ZFF	Zero frequency filtering
ZFFS	Zero frequency filtered signal
ZTW	Zero time windowing
ZTWCCs	Zero time windowing cepstral coefficients

it approximately grows or decays as a polynomial function of time. The growing or decaying trend in $y_o[n]$ is removed by subtracting the local mean computed over the average pitch period at each sample. The resulting signal ($y[n]$) is referred to as the zero frequency filtered signal (ZFFS) and can be expressed as follows:

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

where $2N + 1$ corresponds to the number of samples in the window used for trend removal. The negative-to-positive zero crossings of the ZFFS correspond to the glottal closure instants (GCIs).

2.1.1. Features derived using ZFF

The ZFFS can be regarded as an approximate voice source waveform, and therefore it can be used in the estimation of the voice source characteristics in the time domain [55]. Four features are computed in the current study to quantify the voice source characteristics using the ZFFS at GCIs. These features are the *slope of the ZFFS* (ZFFS slope), the *energy of excitation* (EoE), the *loudness measure* (Loudness) and the *energy of the ZFFS* (ZFFS energy). By denoting GCIs in a voiced segment as $\mathcal{G} = \{g_1, g_2, ..., g_M\}$, where M is the number of GCIs, these four features are computed as follows:

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

$$ZFFS\,slope_{g_c} = |y[g_c + 1] - y[g_c - 1]|, \qquad c = 1, 2, \ldots, M. \tag{3}$$

This feature was used in the analysis and classification of vocal emotions in [58, 59], where the feature was shown to be large for emotions of low arousal (such as sadness), and small for emotions of high arousal (such as anger and happiness). Therefore, this feature reflects changes in the relative duration of the glottal closed phase in a similar manner to CQ and NAQ [16, 60].

EoE is computed from the Hilbert envelope ($he[n]$) of the LP residual of $x[n]$ over a 1-ms region around the c^{th} GCI [59] and is given by:

$$EoE_{g_c} = \frac{1}{2K+1}\sum_{i=-K}^{K} he^2[g_c + i], \qquad c = 1, 2, \ldots, M, \tag{4}$$

where 2K+1 corresponds to the number of samples in the 1-ms window. This feature was shown to reflect the changes in vocal effort [58, 59]. The experiments in [58, 59] indicated that EoE was generally large for emotions of high arousal and small for emotions of low arousal.

Loudness is defined as 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:

$$Loudness_{g_c} = \frac{\sigma_{g_c}}{\mu_{g_c}}, \qquad c = 1, 2, \ldots, M. \tag{5}$$

The loudness measure has been shown to indicate the abruptness of glottal closure [61].

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

$$ZFFS\ energy_{g_c} = \frac{1}{L}\sum_{i=-L/2}^{L/2} y^2[g_c + i], \qquad c = 1, 2, \ldots, M, \tag{6}$$

where L is the window length over which energy is computed. Since the ZFFS is a low-pass filtered signal, a large value of ZFFS energy reflects a prominent low-frequency content of the signal.

2.2. Zero time windowing (ZTW)

ZTW is a frequency domain method to analyse voice source characteristics. The computation of ZTW [56] begins by multiplying the microphone signal in the time domain with a heavily decaying window. The window consists of two parts, $w_1^2[n]w_2[n]$, that are defined as follows:

$$\begin{aligned} w_1[n] &= 0, \quad n = 0, \\ &= \frac{1}{4\sin^2(\pi n/2P)}, \quad n = 1, 2, \ldots, P-1, \end{aligned} \tag{7}$$

$$w_2[n] = 4\cos^2(\pi n/2P), \quad n = 0, 1, \ldots, P-1, \tag{8}$$

where P is the window length in samples. Multiplying a signal with window $w_1^2[n]$ emphasizes the values near the beginning (the zeroth sample) of the signal and hence the name 'zero time windowing'. This time domain multiplication is approximately equivalent to integrating the signal four times in the frequency domain. The numerator of the group delay function (NGD) of the windowed signal (i.e., of $u[n] = w_1^2[n]w_2[n]s[n]$) is computed to estimate the spectrum. NGD is computed as follows:

$$n_{g_d}[k] = \mathrm{Re}\{U[k]\}\mathrm{Re}\{V[k]\} + \mathrm{Im}\{U[k]\}\mathrm{Im}\{U[k]\}, \quad k = 0, 1, 2, \ldots, N-1, \tag{9}$$

where $U[k]$ is the N-point FFT of $u[n]$ and $V[k]$ is the N-point FFT of $v[n]$ ($v[n] = nu[n]$). The NGD function is double-differentiated to emphasize the resonances of the vocal tract system. Furthermore, the low-amplitude peaks in the double-differentiated NGD (DDNGD) are highlighted by computing the Hilbert envelope of DDNGD, and the resulting spectrum is referred to as the ZTW spectrum. The ZTW spectrogram can be obtained by computing the ZTW spectrum at each instant of n, and this is denoted by $S[n, k]$ in this study. More details of the computation of the ZTW spectrum can be found in [56].

The ZTW method provides high temporal resolution with the use of a heavily decaying window being shifted at each sample and simultaneously maintaining a good spectral resolution with the use of group delay. Hence, it is capable of quantifying the time varying characteristics of the voice production mechanism [56]. The ZTW spectrum has been shown to effectively model various voice excitation characteristics, such as the glottal open phase [62, 63], but also time varying vocal tract system characteristics, such as formants [56].

Figs. 1 and 2 show examples of average ZTW spectra and ZTW spectrograms, respectively, for speech signals (the vowel /e/) produced in breathy, modal and pressed phonation types. The ZTW spectra were computed by averaging the ZTW spectrograms over all time instants of the utterance. The spectra show distinguishable differences among the three phonation types. Breathy phonation exhibits a larger low-frequency emphasis in the spectrum compared to modal and pressed. The spectrum of both the modal and pressed utterance shows a prominent peak between 0 Hz and 1000 Hz. This peak is located at a lower frequency (around 400 Hz) in the modal utterance compared to the pressed one (where it is located around 500 Hz). The location of the prominent peak varies with the open phase characteristics of the speech signal. For a longer open phase, the prominent peak in the spectrum shifts to lower frequencies. This is due to the fact that for a longer open phase, the resulting tract (consisting of the subglottal and supraglottal system) is longer, which gives rise to dominant low-frequency characteristics [62, 64, 65]. It is known that breathy voices exhibit longer open phases compared to pressed voices [30]. It can also be observed from Fig. 1 that the energy of high frequencies is larger in pressed phonation compared to modal and breathy, which is in line with previous studies (e.g., [61]).

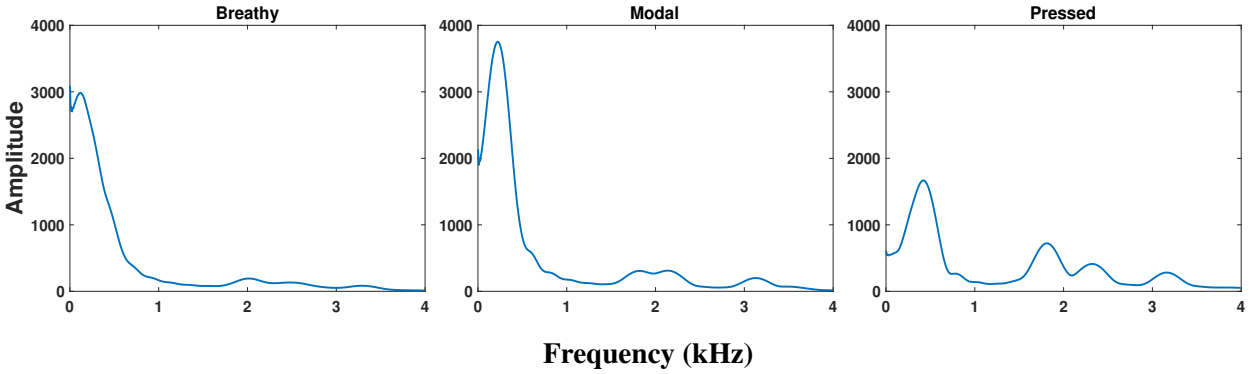

Figure 1: ZTW spectra for breathy, modal and pressed speech signals (the vowel /e/).

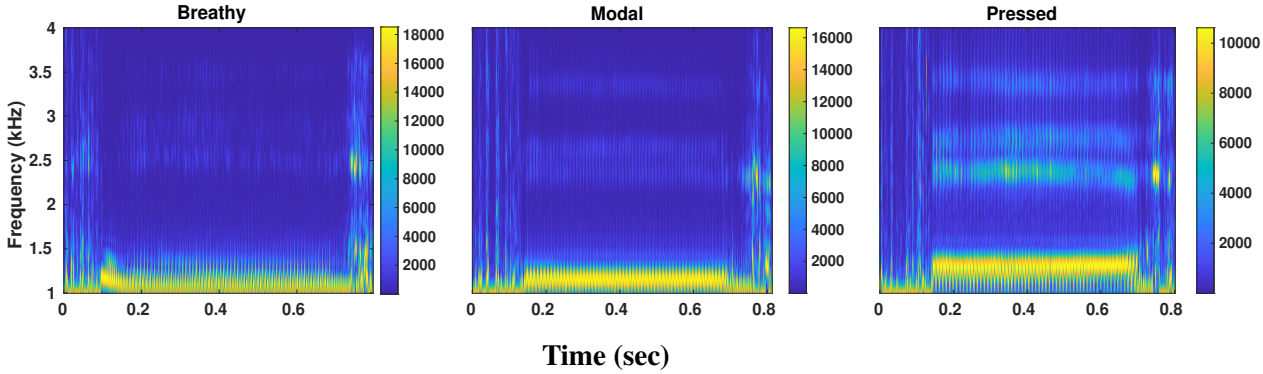

Figure 2: ZTW spectrograms for breathy, modal and pressed speech signals (the vowel /e/).

2.2.1. Features derived using ZTW

In order to parameterize the frequency domain phenomena related to the voice source described in the above examples, five scalar features are first derived using the ZTW spectrum. In these features, spectral moments are used to measure the global spectral shape and the local spectral peak of the spectrum. These five features are the *dominant peak location* (DPL), the *dominant peak amplitude* (DPA), the *centre of gravity* (CoG), the *entropy* (H) and the *band energy ratio* (BER). In addition, as the sixth method, the ZTW spectrum is represented using cepstral coefficients and the corresponding feature vector is referred to as the *zero time windowing cepstral coefficients* (ZTWCCs) [53].

DPL is the location of the dominant peak in the ZTW spectrum. DPL is computed at the c^{th} GCI as follows:

$$DPL_{g_c} = \arg\max_k \{S[g_c, k]\}, \qquad c = 1, 2, \ldots, M, \tag{10}$$

where $S[g_c, k]$ denotes the ZTW spectrum at time instant g_c.

DPA is the amplitude of the largest peak in the ZTW spectrum. DPA is computed at the c^{th} GCI as follows:

$$DPA_{g_c} = \max\{S[g_c, k]\}, \qquad c = 1, 2, \ldots, M. \tag{11}$$

CoG is a measure for the centre of the mass distribution along the frequency axis. CoG indicates tilting of the spectral distribution towards lower or higher frequencies. CoG is computed at the c^{th} GCI as follows:

$$CoG_{g_c} = \frac{\sum_{k=1}^{N} k\, S[g_c, k]}{\sum_{k=1}^{N} S[g_c, k]}, \qquad c = 1, 2, \ldots, M. \tag{12}$$

H is a measure of the average amount of uncertainty of the spectral distribution and it is defined as

$$H_{g_c} = -\sum_{k=1}^{N} \hat{S}[g_c, k] \log_2 \hat{S}[g_c, k], \qquad c = 1, 2, \ldots, M, \tag{13}$$

where

$$\hat{S}[g_c, k] = \frac{S[g_c, k]}{\sum_{k=1}^{N} S[g_c, k]}, \qquad c = 1, 2, \ldots, M. \tag{14}$$

The entropy satisfies the following inequality:

$$0 \leq H \leq \log_2 N. \tag{15}$$

The lower bound in Eq. (15) corresponds to no uncertainty, which occurs when the value of one spectral bin is equal to one and all other bins are zero-valued. The upper bound corresponds to the maximum uncertainty which occurs when all the spectral bins are of an equal value.

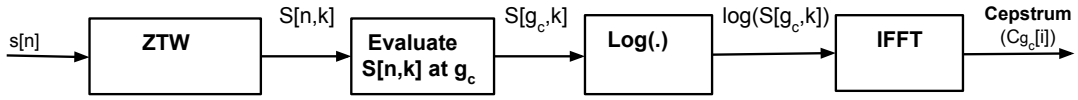

Figure 3: Block diagram of the extraction of zero time windowing cepstral coefficients (ZTWCCs) [53].

BER is the ratio between the high-frequency and low-frequency energy, defined as

$$BER_{g_c} = \frac{\sum_{k=N_c+1}^{N_\pi} S[g_c,k]^2}{\sum_{k=1}^{N_c} S[g_c,k]^2}, \qquad c = 1,2,\ldots,M, \tag{16}$$

where N_c denotes the FFT index of the cut-off frequency (here 1.5 kHz), and N_π corresponds to the Nyquist frequency.

ZTWCCs are obtained by computing the cepstrum using the ZTW spectrum at the c^{th} GCI as follows:

$$C_{g_c}[i] = \text{IFFT}(\log_{10}(S[g_c,k])), \qquad c = 1,2,\ldots,M. \tag{17}$$

From cepstrum $C_{g_c}[i]$, the first 13 cepstral coefficients ($1 \leq i \leq 13$) are considered. In addition to the static coefficients, delta and double-delta coefficients are also included, which makes ZTWCCs 39-dimensional. The schematic block diagram for computing ZTWCCs is shown in Fig. 3.

2.3. Single frequency filtering (SFF)

SFF [57, 66] is a time-frequency analysis technique that can be used to compute an amplitude envelope of the microphone signal as a function of time at a selected frequency. The amplitude envelope is obtained by first frequency-shifting (i.e., modulating) the microphone signal $s[n]$ by multiplying it with an exponential function as follows: $\hat{s}[n,k] = s[n]e^{-j2\pi\bar{f}_k n/f_s}$, where f_s is the sampling frequency, $\bar{f}_k = \frac{f_s}{2} - f_k$, and f_k is the k^{th} desired frequency. The modulated signal is filtered using a single-pole filter, whose transfer function is $H(z) = \frac{1}{1+rz^{-1}}$. The pole of the filter ($z = -r$) is located on the negative real axis close to the unit circle. In this study, we chose $r = 0.995$. The output of the single-pole filter is given by

$$y[n,k] = -ry[n-1,k] + \hat{s}[n,k]. \tag{18}$$

The amplitude envelope ($v[n,k]$) of $y[n,k]$ at frequency f_k is given by

$$v[n,k] = \sqrt{(y_r[n,k])^2 + (y_i[n,k])^2}, \tag{19}$$

where $y_r[n,k]$ and $y_i[n,k]$ correspond to the real and imaginary parts of $y[n,k]$, respectively. The amplitude envelope of the microphone signal can be computed for several frequencies at intervals of Δf by defining f_k as follows:

$$f_k = k\Delta f, \qquad k = 1,2,\ldots,K, \tag{20}$$

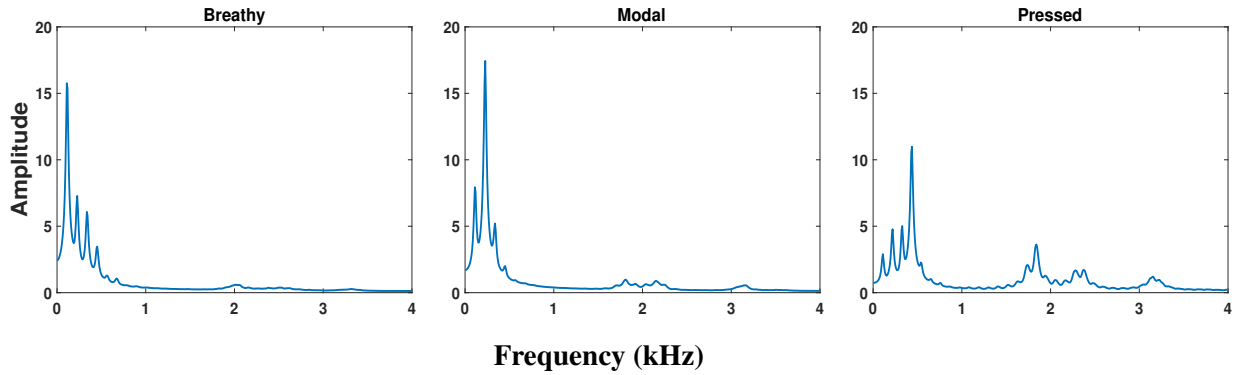

Figure 4: SFF spectra for breathy, modal and pressed speech signals (the vowel /e/).

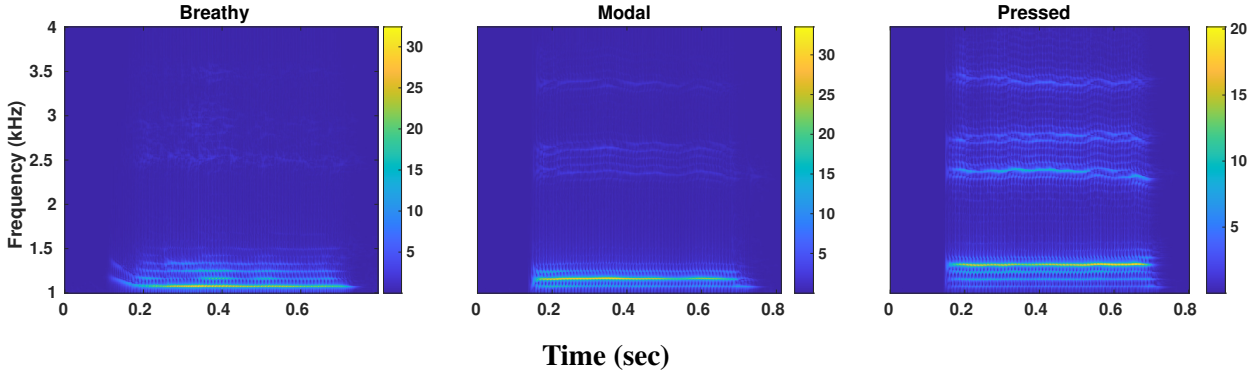

Figure 5: SFF spectrograms for breathy, modal and pressed speech signals (the vowel /e/).

where $K = \frac{(f_s/2)}{\Delta f}$. In this study, we chose Δf=10 Hz. The SFF magnitude spectrum can be obtained for each instant of time from $v[n, k]$.

Figs. 4 and 5 show examples of average SFF spectra and SFF spectrograms, respectively, for speech signals (the vowel /e/) produced in breathy, modal and pressed phonation types. The SFF spectra were computed by averaging the SFF spectrograms over all time instants of the utterance. The breathy voice shows a more prominent first harmonic compared to the modal and pressed utterances. The pressed voice shows a dominant fourth spectral harmonic, which is beyond the dominant harmonic (the second harmonic) in the spectrum of the modal utterance. This is in line with previous studies reported in [28]. It can also be observed that the spectrum of the pressed voice appears to be flatter than the spectra of the modal and breathy voices. As previously reported in [61], this is most likely due to sharper glottal closures in pressed phonation.

2.3.1. Features derived using SFF

The spectral examples above demonstrate that the SFF spectrum is associated with the sharpness of the glottal closure. In order to parameterize the spectrum, three scalar features based on a previous study [66, 63] are used. These three features are the *slope of spectral gain* (SSG), the *slope of spectral variance* (SSV) and the *spectral flatness* (SF). In addition, cepstral coefficients are computed from the SFF spectrum, and they are referred to as single frequency filtering cep-

stral coefficients (SFFCCs) [54]. Moreover, the five scalar features (DPL, DPA, CoG, entropy and BER) that were described in Section 2.2.1 related to the ZTW spectrum are computed from the SFF spectrum.

SSG is based on the fact that glottal closure is a high-energy event in the time domain. Therefore, the sum (i.e., the spectral gain) of the amplitude envelope $v[n,k]$ reaches local maximum at glottal closure. Spectral gain (SG) is computed as follows:

$$SG[n] = \frac{1}{K}\sum_{k=1}^{K} v[n,k], \qquad n = 0, 1, \ldots, N_s - 1, \tag{21}$$

where N_s corresponds to the number of samples of the microphone signal. The slope of SG (SSG) at the c^{th} GCI is computed as follows:

$$SSG_{g_c} = |SG[g_c + 1] - SG[g_c - 1]|, \qquad c = 1, 2, \ldots, M. \tag{22}$$

SSV is based on the fact that an impulse-like excitation in the time domain results in a flat spectrum with a low variance. Spectral variance (SV) is computed as follows:

$$SV[n] = \frac{1}{K}\sum_{k=1}^{K} (\hat{v}[n,k] - \mu[n,k])^2, \qquad n = 0, 1, \ldots, N_s - 1, \tag{23}$$

where

$$\hat{v}[n,k] = \frac{v[n,k]}{\sum\limits_{k=1}^{K} v[n,k]}, \qquad n = 0, 1, \ldots, N_s - 1, \tag{24}$$

$$\mu[n,k] = \frac{1}{K}\sum_{k=1}^{K} \hat{v}[n,k] = \frac{1}{K}, \qquad n = 0, 1, \ldots, N_s - 1. \tag{25}$$

Due to normalization, the amplitude information at each sample is lost. The slope of SV (SSV) at the c^{th} GCI is computed as follows:

$$SSV_{g_c} = |SV[g_c + 1] - SV[g_c - 1]|, \qquad c = 1, 2, \ldots, M. \tag{26}$$

SF is the ratio between the geometric and arithmetic means [67]. It is justified to be used as a feature to characterize the sharpness of glottal closure because a sharp closure results generally in a flat spectrum. SF at the c^{th} GCI is computed as follows:

$$SF_{g_c} = \frac{\sqrt[K]{\prod\limits_{k=1}^{K} \hat{v}[g_c,k]}}{\frac{1}{K}\sum\limits_{k=1}^{K} \hat{v}[g_c,k]}, \qquad c = 1, 2, \ldots, M. \tag{27}$$

SF is always between 0 and 1. For a perfectly flat spectrum, the value of SF is 1.

SFFCCs are computed as the cepstrum from the SFF spectrum at the c^{th} GCI as follows:

$$C_{g_c}[i] = \text{IFFT}(\log_{10}(v[g_c, k])), \qquad c = 1, 2, \ldots, M. \tag{28}$$

From $C_{g_c}[i]$, the first 13 cepstral coefficients ($1 \leq i \leq 13$) are considered. In addition to the static coefficients, delta and double-delta coefficients are also included, which makes SFFCCs 39-dimensional. The schematic block diagram describing the computation of SFFCCs is shown in Fig. 6.

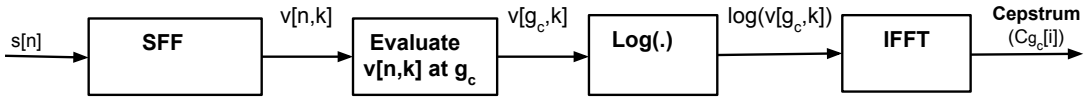

Figure 6: Block diagram of the extraction of single frequency filtering cepstral coefficients (SFFCCs) [54, 68].

3. Analysis of Phonation Types in Speech and Singing Voice

In this section, phonation types are analysed both from speech and singing voice using the features described in Section 2. Feature distributions (Figs. 7–12) are depicted between phonation types for all the features using box plots, where the central mark indicates the median, and the bottom and top edges indicate the 25^{th} and 75^{th} percentiles, respectively. Whiskers describe all points within 1.5 times the interquartile range, and points beyond these whiskers are plotted as outliers using the $'+'$ symbol. In addition to depicting feature distributions, statistical tests are carried out using 1-way ANOVAs to analyse how different features are capable of separating phonation types.

3.1. Databases

3.1.1. Phonation type database of speech

The database used in the current investigation to study phonation types in speech consists of eight different Finnish vowels uttered in three phonation types (breathy, modal and pressed) by six female and five male speakers, aged between 18 and 48 years. Each vowel was uttered three times, resulting in a total of $3 \cdot 3 \cdot 8 \cdot 11 = 792$ isolated vowels. The database was originally recorded at a sampling frequency of 44.1 kHz in an anechoic chamber and later downsampled to 16 kHz. More details of the database can be found in [16, 69].

3.1.2. Phonation type database of singing voice

The singing voice database used contains sustained vowels sung by a professional Russian soprano female singer [39]. The phonation types include breathy, neutral/modal, flow and tense/pressed voice [41]. The database consists of 763 voice signals, including nine different vowels whose pitch ranges from A3 to G5. In this study, voices produced using flow phonation are not considered because they do not correspond to Sundberg's definition of flow phonation [41, 70]. The data were originally recorded using a sampling frequency of 96 kHz. More details of the database can be found in [39].

In this study, both of the databases are downsampled to 8 kHz for feature extraction. It is to be noted that even though the absolute feature values for some of the features may vary with the sampling frequency, the trend with reference to phonation type (from breathy to modal and then to pressed) will not change.

Both of the databases described above are much smaller than databases that are used currently in speech technology areas such as speech recognition and speech synthesis. However, to the best of our knowledge, these two corpora are the only phonation type databases that are publicly available currently for research purposes. We would also like to point out that acquiring data with reliable ground truth (i.e., phonation type label) is not simple because phonation types occur on a continuum in natural production of speech and singing voice. Ideally, each recorded voice sample should be evaluated by a panel of expert listeners to ensure that the sample represents perceptually the desired phonation type.

3.2. Feature Analysis in Speech

The distributions of the features derived using the ZFF, ZTW and SFF methods are shown in Figs. 7, 8 and 9, respectively, for the three phonation types (breathy, modal and pressed) of the speech database described in Section 3.1.1.

Fig. 7 shows the distributions of the features derived using the ZFF method. It can be seen that both ZFFS slope and ZFFS energy show a decreasing trend when the phonation type changes from breathy to modal and then to pressed. For ZFFS slope, this trend reflects the increase of the duration of the glottal closed phase. The decreasing trend in ZFFS energy in turn depicts the reduction of the low-frequency contents of the voice source spectrum when the phonation type changes from breathy to modal and then to pressed. EoE and Loudness follow an increasing trend. The trend in EoE depicts the increased vocal effort and the trend in loudness describes the increased sharpness of glottal closure when moving from breathy to modal and then to pressed.

Fig. 8 shows the distributions of the features derived from the ZTW method. It can be seen that all the features except DPA show an increasing trend when the phonation type changes from breathy to modal and then to pressed. This increasing trend is caused by the reduction of the duration of the glottal open phase as a function of phonation type. For a longer open phase of the glottal flow, the spectral energy of the glottal source concentrates mainly on lower frequencies, whereas for a shorter open phase the energy is more spread to higher frequencies. For the same reason, DPA shows a decreasing trend when the phonation type changes from breathy to modal and then to pressed.

Fig. 9 depicts the distributions of the features computed using the SFF method. From the eight SFF features, five (SF, DPL, CoG, H, BER) show an increasing trend when the phonation type changes from breathy to modal and then to pressed. This increasing trend is caused by the increase of the duration of the glottal closed phase as a function of phonation type. For a longer closed phase of the glottal flow, the spectral energy of the glottal source spreads to higher frequencies, whereas for a shorter closed phase the energy concentrates mainly on lower frequencies. Three of the features (SSG, SSV, DPA) follow a decreasing trend reflecting the changes in glottal closed phase, similar to ZFFS slope.

We would like to point out that from the view of their technical definition, the proposed features should not depend on issues such as intensity, pitch and vowel identity. Instead, the features should

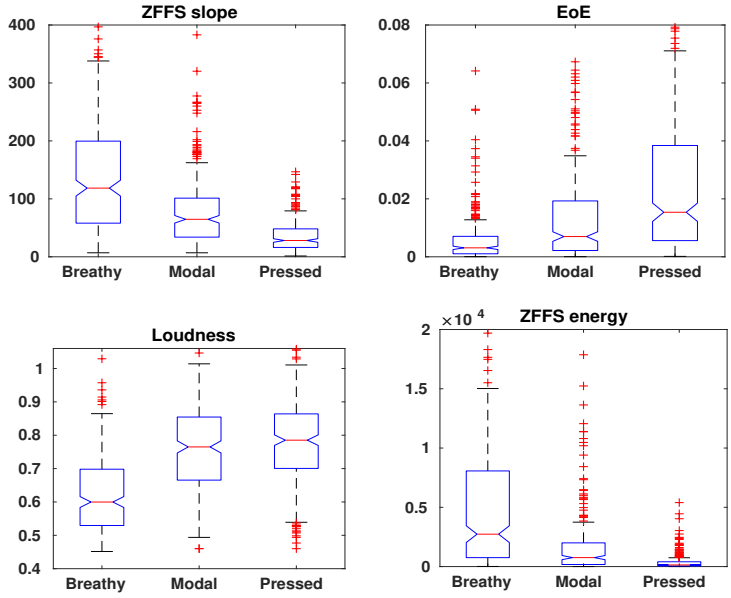

Figure 7: Distribution of features computed using the ZFF method for breathy, modal and pressed phonation types in speech.

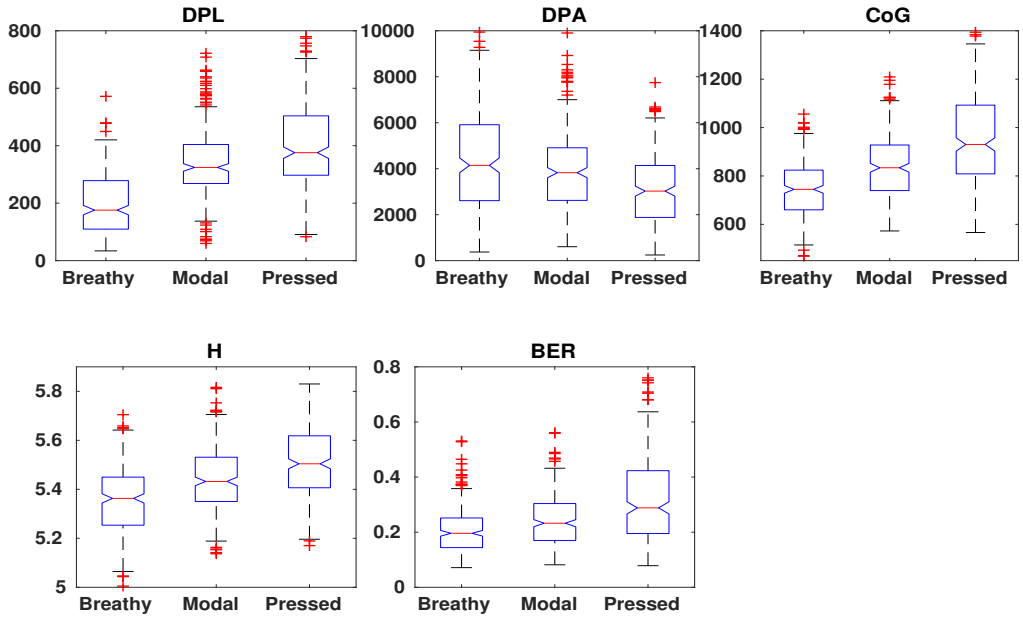

Figure 8: Distribution of features computed using the ZTW method for breathy, modal and pressed phonation types in speech.

reflect changes that occur in phonation type, which is evident from the above results. However, one should not interpret this in such a manner that the developed features are not affected at all by, for example, changes in pitch. In natural production of speech and singing voice, pitch and phonation type are interrelated, which might result in dependencies between the proposed features

and pitch. This phenomenon is, however, outside the scope of the present study.

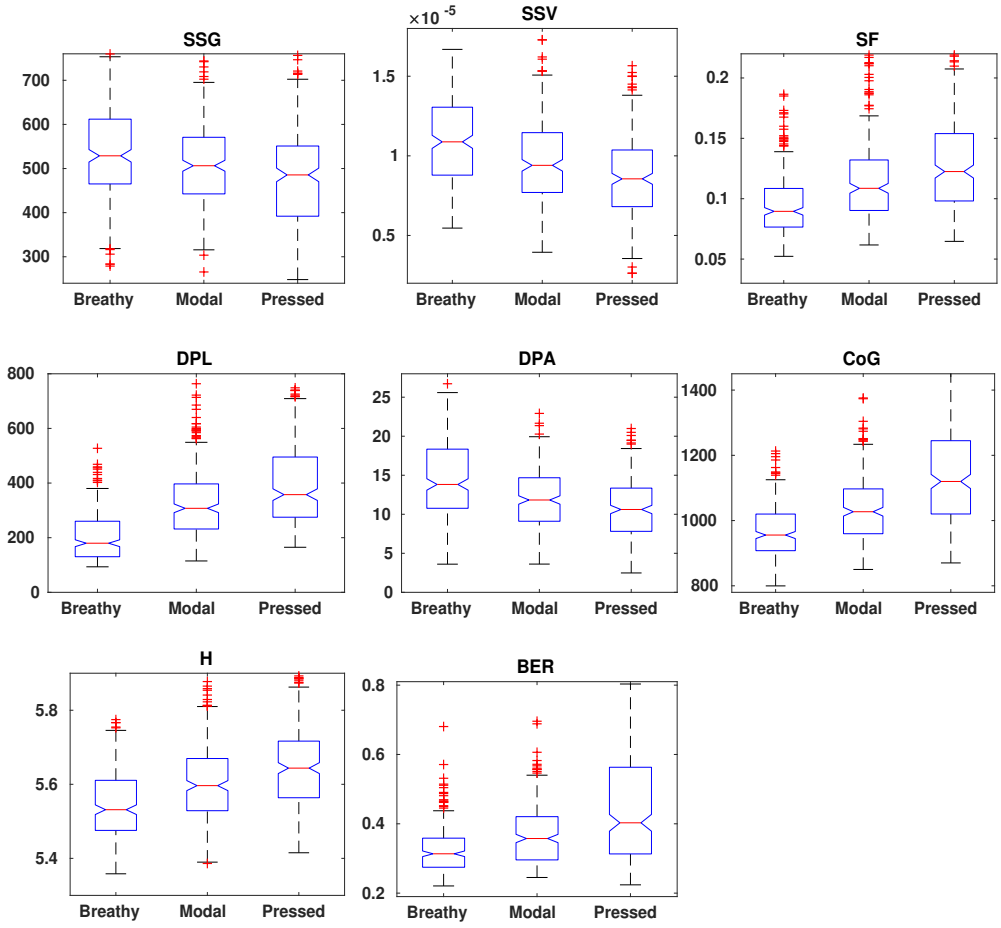

Figure 9: Distribution of features computed using the SFF method for breathy, modal and pressed phonation types in speech.

In order to analyse whether the different features are capable of yielding statistically significant differences between the three phonation types, one-way ANOVAs were computed for the features derived using the ZFF, ZTW and SFF methods. In these statistical analyses, feature was the dependent variable, and phonation type (breathy, modal and pressed) was the independent variable (i.e., the number of degrees of freedom was 2). Furthermore, multiple comparisons of the phonation types with regard to each feature were carried out using Tukey's honestly significant difference (HSD) test. The results of the ANOVA tests are given in Table 2. From the table, it can be observed that all the features showed statistically significant ($p < 0.001$) differences between the phonation types. Table 3 shows the results of the multiple comparisons tests. It can be seen that for all the proposed features, there were statistically significant differences between the phonation types studied except between breathy and modal with the SSG feature (computed using the SFF method) and between modal and pressed with the loudness feature (computed using the ZFF method).

Table 2: One-way ANOVA results for the features derived from the ZFF, ZTW and SFF methods for phonation types in speech (the number of degrees of freedom is 2). SS - Sum square, MS - Mean square, χ^2 - Chi-Square, F - F value, p - probability.

Method	**SS**	**MS**	χ^2	**F**	**p**
ZFF					
ZFFS slope	$16x10^5$	$7.9x10^5$	0.27	147.4	< 0.001
EoE	0.05	0.03	0.17	077.4	< 0.001
Loudness	3.41	1.71	0.22	107.7	< 0.001
ZFFS energy	$8.1x10^9$	$4.1x10^9$	0.14	059.4	< 0.001
ZTW					
DPL	$6.6x10^6$	$3.3x10^6$	0.31	171.1	< 0.001
DPA	$2.7x10^8$	$1.3x10^8$	0.09	038.8	< 0.001
CoG	$56x10^5$	$28x10^5$	0.25	131.6	< 0.001
H	3.14	1.57	0.18	082.9	< 0.001
BER	2.19	1.09	0.15	065.9	< 0.001
SFF					
SSG	$5.1x10^5$	$2.5x10^5$	0.05	018.1	< 0.001
SSV	$7.1x10^{-9}$	$3.5x10^{-9}$	0.12	049.6	< 0.001
SF	0.22	0.11	0.14	060.7	< 0.001
DPL	$4.9x10^6$	$2.5x10^6$	0.27	146.7	< 0.001
DPA	$1.9x10^3$	$9.7x10^2$	0.12	052.2	< 0.001
CoG	$4.3x10^6$	$2.1x10^6$	0.28	155.2	< 0.001
H	1.45	0.73	0.15	070.5	< 0.001
BER	2.76	1.38	0.17	081.5	< 0.001

Table 3: Multiple comparison of phonation types in speech with regard to all proposed features. Symbol $*$ indicates a significant difference ($p < 0.05$) between phonation types.

	Loudness		**SSG**		**Remaining features**	
	Modal	Pressed	Modal	Pressed	Modal	Pressed
Breathy	*	*	–	*	*	*
Modal		–		*		*

3.3. Feature Analysis in Singing Voice

Distributions of the features computed from singing voices using the ZFF, ZTW and SFF methods are shown in Figs. 10, 11 and 12, respectively. These figures depict feature distributions as a function of phonation type for the three phonation types (breathy, modal and pressed) of the singing voice database described in Section 3.1.2. From the figures, it can be observed that all the individual features extracted from singing voice follow a increasing or decreasing trend as a function of phonation type, similar to the features' trend showed for speech in Section 3.2.

To analyse the statistical significance of the features, ANOVAs were carried out and the results

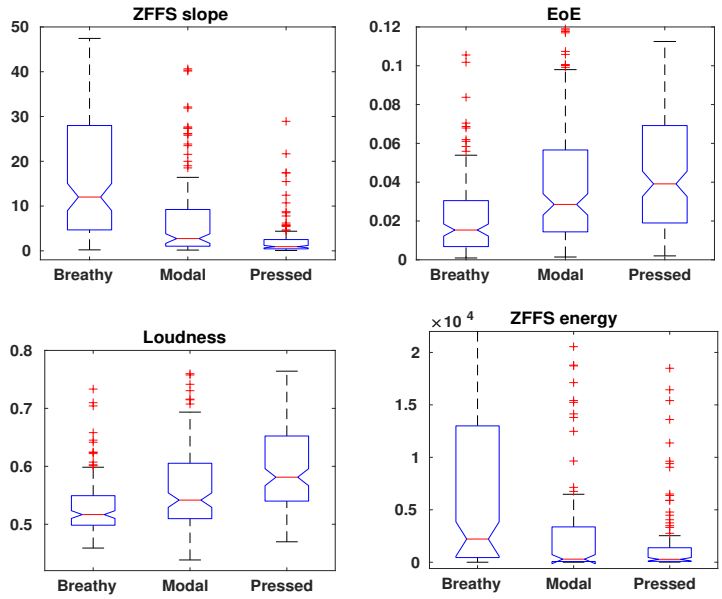

Figure 10: Distribution of features computed using the ZFF method for breathy, modal and pressed phonation types in singing.

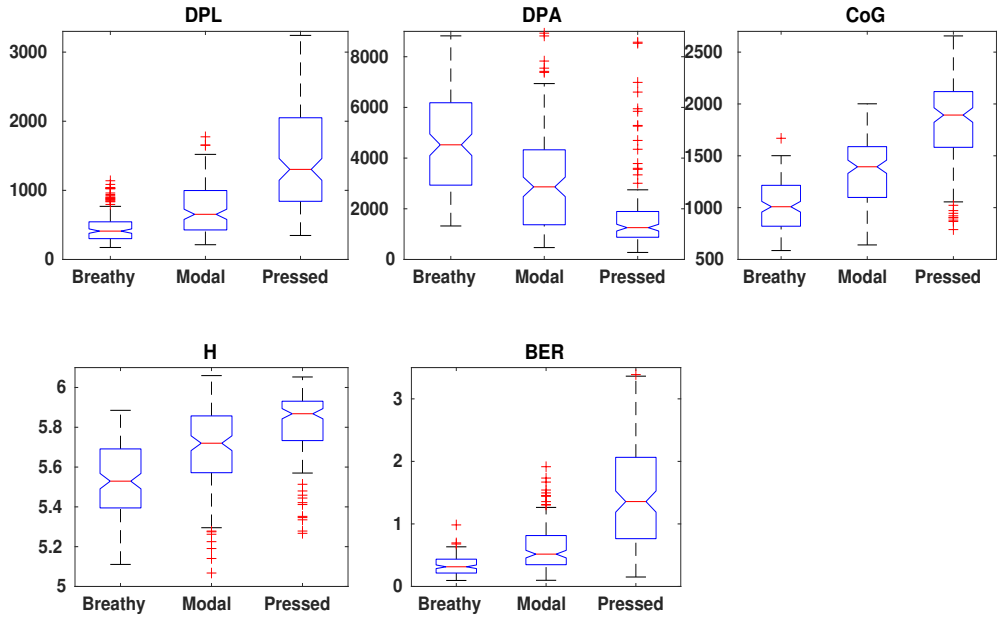

Figure 11: Distribution of features computed using the ZTW method for breathy, modal and pressed phonation types in singing.

are reported in Table 4. The feature was again the dependent variable, and the phonation type (breathy, modal and pressed) was the independent variable (i.e., the number of degrees of freedom was 2). From the table, it can be observed that all the features showed statistical significance ($p < 0.001$). A multiple comparison of different phonation types with regard to each feature

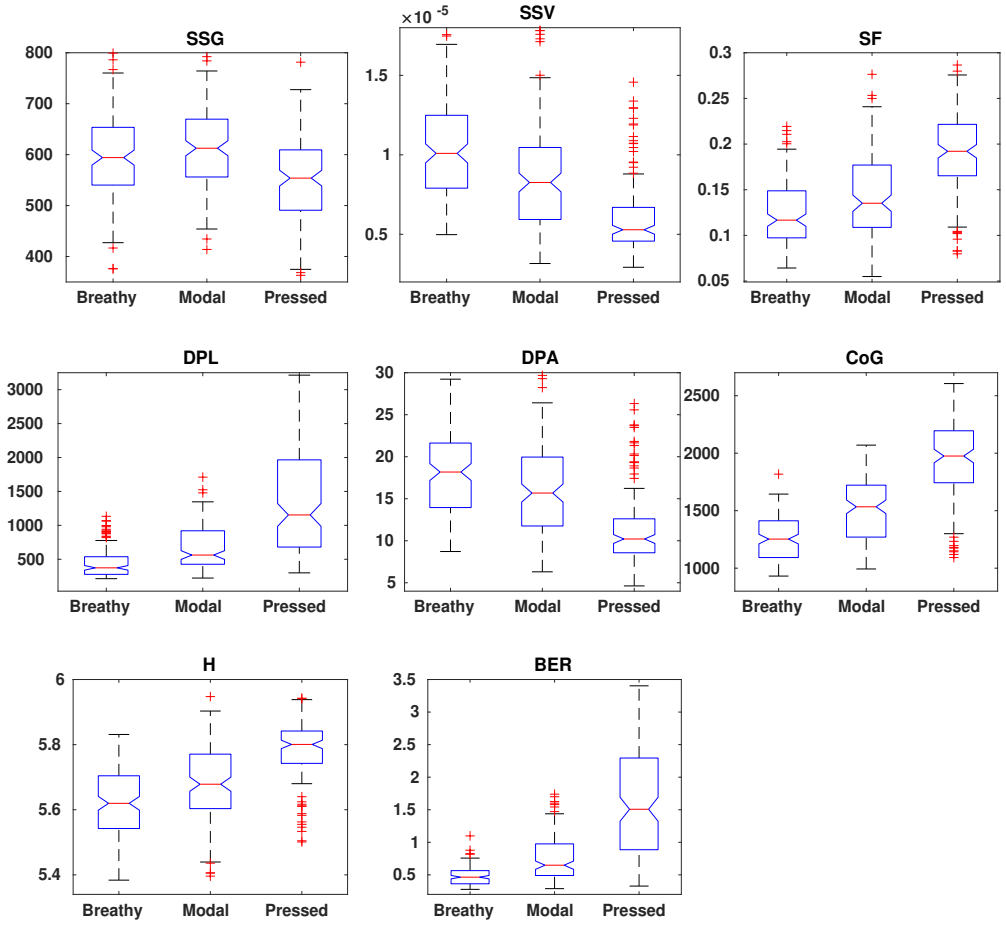

Figure 12: Distribution of features computed using the SFF method for breathy, modal and pressed phonation types in singing.

was carried out with Tukey's HSD test and the results are given in Table 5. From the table, it can be observed that for all the proposed features, there were statistically significant differences between the phonation types except between modal and pressed with the EoE and ZFFS energy features (computed using the ZFF method) and between breathy and modal with the SSG feature (computed using the SFF method).

4. Classification of Phonation Types in Speech and Singing Voice

In order to study how phonation types can be automatically classified with machine learning, the proposed features described in Section 2 were used to train and evaluate a classifier. In addition, the proposed features and baseline features were combined in classification experiments. This section describes the feature sets used in the classification experiments, the details of the classifier and the results of the classification experiments for phonation types in speech and singing voice.

Table 4: One-way ANOVA results for the features computed using the ZFF, ZTW and SFF methods for phonation types in singing (the number of degrees of freedom is 2). SS - Sum square, MS - Mean square, χ^2 - Chi-Square, F - F value, p - probability.

Method	**SS**	**MS**	χ^2	**F**	**p**
ZFF					
ZFFS slope	$2.5x10^4$	$1.2x10^4$	0.19	152.7	< 0.001
EoE	0.07	0.03	0.12	080.3	< 0.001
Loudness	0.22	0.11	0.08	054.9	< 0.001
ZFFS energy	$6.9x10^9$	$3.4x10^9$	0.11	079.3	< 0.001
ZTW					
DPL	$1.1x10^8$	$0.6x10^8$	0.38	372.1	< 0.001
DPA	$1.3x10^9$	$0.7x10^9$	0.35	332.6	< 0.001
CoG	$7.6x10^7$	$3.8x10^7$	0.52	657.5	< 0.001
H	10.51	05.25	0.34	316.0	< 0.001
BER	$1.9x10^2$	95.82	0.33	303.3	< 0.001
SFF					
SSG	$4.2x10^5$	$2.1x10^5$	0.09	062.3	< 0.001
SSV	$2.1x10^{-9}$	$1.1x10^{-9}$	0.29	258.8	< 0.001
SF	0.42	0.21	0.26	221.6	< 0.001
DPL	$9.1x10^7$	$4.5x10^7$	0.35	325.7	< 0.001
DPA	$6.8x10^3$	$3.4x10^3$	0.31	274.9	< 0.001
CoG	$5.1x10^7$	$2.5x10^7$	0.51	634.5	< 0.001
H	2.63	1.32	0.29	254.3	< 0.001
BER	$1.8x10^2$	91.28	0.36	340.2	< 0.001

Table 5: Multiple comparison of phonation types in singing with regard to all proposed features. Symbol $*$ indicates a significant difference ($p < 0.05$) between phonation types.

	EoE and ZFFS energy		**SSG**		**Remaining features**	
	Modal	Pressed	Modal	Pressed	Modal	Pressed
Breathy	*	*	–	*	*	*
Modal		–		*		*

4.1. Feature Sets

The proposed features were grouped into five individual feature sets. Three of these feature sets consist of the scalar features computed using the three signal processing methods (ZFF, ZTW, and SFF) described in Section 2. These three sets are referred to as the ZFF feature set (consisting of ZFFS slope, EoE, Loudness, and ZFFS energy), the ZTW feature set (consisting of DPL, DPA, CoG, H, and BER), and the SFF feature set (consisting of SSG, SSV, SF, DPL, DPA, CoG, H, and BER). The remaining two feature sets include cepstral coefficients, ZTWCCs and SFFCCs described in Sections 2.2.1 and 2.3.1, respectively.

Glottal source features (referred to here as the voice quality (VQ) features as in [3]) and MFCCs were chosen as baseline features for comparison. The selection of these features is based on the results of [3, 16, 30], which indicate that these features gave the best performance in the discrimination of phonation types. The VQ feature set consists of NAQ [60], QOQ [16], H1-H2 [31], PSP [34] and MDQ [3]. The first four of these features were computed using iterative adaptive inverse filtering [71] as the GIF method, and MDQ was derived by computing the wavelet decomposition from the LP residual [3]. All the features were computed using the COVAREP toolbox [72]. Conventional 13-dimensional MFCCs were computed using 25-ms Hamming-windowed frames with a 5-ms frame shift. In addition to the static coefficients, delta and double-delta coefficients were also computed, resulting in a 39-dimensional feature vector.

The feature sets described above were also combined in order to study complementary information among the features. In total, 14 feature sets (FSs) were created as listed below. From these 14 feature sets, the first seven (from FS-1 to FS-7) consist of the individual feature sets described above. The rest of the feature sets (from FS-8 to FS-14) were combined from the existing features sets (FS-8) and from the proposed feature sets (FS-9 to FS-13). The last of the combined set (FS-14) was built by combining the existing feature set that showed the highest classification accuracy with the proposed set that yielded the highest accuracy. In other words, FS-14 included the best set of the three existing feature sets (FS-1, FS-2 and FS-8) combined with the best set of the ten sets consisting of the proposed features (from FS-3 to FS-7 and from FS-9 to FS-13). The 14 feature sets are:

(i). FS-1: VQ feature set

(ii). FS-2: MFCCs

(iii). FS-3: ZFF feature set

(iv). FS-4: ZTW feature set

(v). FS-5: SFF feature set

(vi). FS-6: ZTWCCs

(vii). FS-7: SFFCCs

(viii). FS-8: VQ feature set+MFCCs

(ix). FS-9: ZTW feature set+ZTWCCs

(x). FS-10: SFF feature set+SFFCCs

(xi). FS-11: ZFF feature set+ZTW feature set+ZTWCCs

(xii). FS-12: ZFF feature set+SFF feature set+SFFCCs

(xiii). FS-13: ZFF feature set+ZTW feature set+ZTWCCs+SFF feature set +SFFCCs

(xiv). FS-14: Combination of the best existing feature set (i.e., the best of FS-1, FS-2 and FS-8) and the best proposed feature set (i.e., the best of FS-3 to FS-7 and FS-9 to FS-13).

4.2. Classifier

Classification experiments were carried out using support vector machines (SVMs) utilizing a radial basis function kernel [73]. The SVM classifier was selected because it is known to be an effective classifier, particularly in cases like the current study where only a small amount of training data is available [3, 32]. Experiments were conducted using 10-fold cross-validation, where data was randomly partitioned into 10 equal sets. One set was held out for testing and the remaining nine sets for training. Classification accuracies were saved for each fold, and finally the mean and standard deviation of the accuracies were computed.

4.3. Results

This section reports the results of the phonation type classification experiments separately in speech and singing.

4.3.1. Classification Results for Phonation Types in Speech

Results from the 10-fold cross-validation experiment for phonations types in speech are shown in terms of the mean and standard deviation of the classification accuracy in Table 6. From the table, it can be observed that the proposed ZTWCCs (FS-6) and SFFCCs (FS-7) gave the highest performance for the individual feature sets. Among the existing feature sets, MFCCs (FS-2) showed better performance than the VQ features (FS-1). Even though the other individual feature sets (FS-3 to FS-5) gave lower performance, they showed complementary information when combined with the proposed features, as shown by the accuracies obtained for FS-9 to FS-13. It can also be observed that there exists complementary information between the existing features: the performance for the combination of the VQ features and MFCCs (FS-8) was higher than that of the corresponding individual feature sets. Among the combinations of the proposed feature sets, the best performance was achieved by FS-13. Hence, FS-14 was selected to consist of FS-8 and FS-13. It can be observed that this feature set resulted in improved accuracy, indicating that there is complementary information between the proposed features and the existing features.

Tables 7, 8 and 9 show confusion matrices for the combination of the existing feature sets (FS-8), for the combination of the proposed feature sets (FS-13) and for the combination of the existing and proposed feature sets (FS-14), respectively. The confusion matrix for FS-8 (Table 7) shows good accuracy for breathy phonation, but confusions between modal and pressed speech signals. This observation is in line with the results reported in [3, 30]. The accuracy shown in Table 8 is better than in Table 7, but still breathy and pressed voices show confusions with modal voices. Compared to the results achieved with feature sets FS-8 and FS-13, classification accuracy shown for FS-14 in Table 9 is remarkably improved. In this case, it can be observed that for all the phonation types, the classification accuracy is improved.

4.3.2. Classification Results for Phonation Types in Singing Voice

The classification results for phonation types in singing are shown in Table 10. In the case of individual feature sets, the performance of the proposed ZTWCCs (FS-6) and SFFCCs (FS-7) is comparable or better than that of MFCCs (FS-2). It is to be noted that the discrimination of phonation types using the VQ features (FS-1) is slightly worse than in speech. However, as in speech, there is complementary information between the VQ features and MFCCs (i.e., FS-8).

Table 6: Mean and standard deviation of the accuracy in phonation type classification of speech for individual feature sets and combinations of feature sets.

Feature set	**Mean**[%]	**Standard deviation**[%]
FS-1	64.21	4.97
FS-2	68.52	5.14
FS-3	61.26	5.84
FS-4	52.62	5.19
FS-5	54.68	5.49
FS-6	69.38	4.53
FS-7	70.88	3.70
FS-8	73.79	4.31
FS-9	69.11	3.17
FS-10	75.44	4.27
FS-11	72.28	3.84
FS-12	75.32	5.03
FS-13	75.06	3.72
FS-14	**78.71**	3.58

Table 7: Confusion matrix in phonation type classification of speech with FS-8.

	Breathy [%]	Modal [%]	Pressed [%]
Breathy	**80.46**	18.01	1.53
Modal	13.67	**66.02**	20.31
Pressed	2.56	22.71	**74.73**

Table 8: Confusion matrix in phonation type classification of speech with FS-13.

	Breathy [%]	Modal [%]	Pressed [%]
Breathy	**79.41**	16.10	4.49
Modal	16.98	**66.80**	16.22
Pressed	2.27	18.97	**78.79**

Table 9: Confusion matrix in phonation type classification of speech with FS-14.

	Breathy [%]	Modal [%]	Pressed [%]
Breathy	**83.16**	15.05	1.79
Modal	14.18	**69.70**	16.12
Pressed	0.84	15.90	**83.26**

Even though the other individual sets consisting of the proposed features (i.e., FS-3 to FS-5) show low performance, they provide complementary information when combined with ZTWCCs and SFFCCs (i.e., FS-9 to FS-13). Further, the combination of all the proposed feature sets (FS-13) resulted in better performance compared to individual feature sets alone (FS-3 to FS-7) and combinations represented by the sets from FS-9 to FS-12. This indicates that there is complementary information among the proposed features. Hence, the final feature set FS-14 was selected to include FS-8 (the best existing feature set) and FS-13 (the best proposed feature set). As in speech, the best classification accuracy was achieved with FS-14, indicating complementary information between the proposed features, and the existing VQ features and MFCCs feature.

Table 10: Mean and standard deviation of the accuracy in phonation type classification of singing voice for individual feature sets and combinations of feature sets.

Feature set	**Mean**[%]	**Standard deviation**[%]
FS-1	60.82	6.61
FS-2	77.07	5.09
FS-3	62.58	5.25
FS-4	65.34	5.26
FS-5	66.71	4.89
FS-6	77.66	4.03
FS-7	78.84	5.15
FS-8	78.95	4.72
FS-9	79.92	6.11
FS-10	80.17	3.94
FS-11	80.16	4.83
FS-12	80.21	4.72
FS-13	82.58	4.35
FS-14	**85.24**	4.80

Table 11: Confusion matrix in phonation type classification of singing voice with FS-8.

	Breathy [%]	Modal [%]	Pressed [%]
Breathy	**75.27**	21.27	03.46
Modal	15.03	**63.34**	19.93
Pressed	01.12	09.11	**89.77**

Tables 11, 12 and 13 show confusion matrices when the classification was based on combination of the existing feature sets, i.e., the VQ features and MFCCs (FS-8), on the combination of all the proposed feature sets (FS-13), and on the combination of existing features and the proposed feature sets (i.e., FS-14), respectively. From Table 11, it can be seen that there are clear confusions between breathy and modal voices, and that modal voices are confused with pressed voices. This

Table 12: Confusion matrix in phonation type classification of singing voice with FS-13.

	Breathy [%]	Modal [%]	Pressed [%]
Breathy	**78.08**	18.02	03.90
Modal	17.56	**66.96**	15.48
Pressed	02.80	02.80	**94.40**

Table 13: Confusion matrix in phonation type classification of singing voice with FS-14.

	Breathy [%]	Modal [%]	Pressed [%]
Breathy	**84.82**	12.89	02.29
Modal	17.26	**72.62**	10.12
Pressed	02.52	04.33	**93.15**

observation is in line with the results reported in [39, 51, 47]. Compared to Table 11, the classification accuracy shown in Table 12 (FS-13) is better for all the three phonation types. Compared to the results achieved with feature sets FS-8 and FS-13, classification accuracy obtained with FS-14 in Table 13 is higher, and the discrimination between breathy and modal voices as well as between modal and pressed voices is further improved. It should be noted that even though there is an improvement, there are still confusions between breathy and modal voices, and modal voices are confused with pressed voices. Further investigations are required to develop features which better reflect differences in voice production characteristics between these classes.

5. Conclusions

In this article, analysis and classification of phonation types was studied in speech and singing voice. Using three signal processing methods (ZFF, ZTW and SFF), several features that reflect changes in the glottal source when phonation type is altered were derived. The proposed features were grouped into five individual feature sets. Three of these feature sets consist of the scalar features computed using the three signal processing methods–they were referred to as the ZFF feature set (ZFFS slope, EoE, Loudness, and ZFFS energy), the ZTW feature set (DPL, DPA, CoG, H, and BER) and the SFF feature set (SSG, SSV, SF, DPL, DPA, CoG, H, and BER). The remaining two feature sets include cepstral coefficients computed from the ZTW and SFF spectra, referred to as ZTWCCs and SFFCCs, respectively. All the proposed features can be computed from the microphone signal without computing the source-filter decomposition.

Statistical analyses were computed in order to study how the ZFF feature set, the ZTW feature set and the SFF feature set are affected when the phonation type in speech and singing voice changes. Multiple comparisons with Tukey's honestly significant difference test showed that for speech, all the features indicated statistically significant differences between the phonation types except SSG (between breathy and modal) and loudness (between modal and pressed). Similarly, for the phonation types in singing, all the features indicated statistically significant differences

between the phonation types except EoE and ZFFS energy (between modal and pressed) and SSG (between breathy and modal).

Classification experiments with SVM revealed that, among the proposed five feature sets, the highest classification accuracy was obtained by ZTWCCs and SFFCCs. The classification accuracy was, however, improved when individual feature sets were combined. The best performance was achieved when all the proposed feature sets were combined. In addition, it was observed that the proposed features provide complementary information to the existing voice quality features and MFCCs that improved the classification of phonation types.

6. Acknowledgements

The first author would like to thank the Academy of Finland (Project No. 312490) for supporting his stay in Finland as a Postdoctoral Researcher. The third author would like to thank the Indian National Science Academy (INSA) for their support.

References

[1] Ingo R Titze. Principles of voice production (second printing). *Iowa City, IA: National Center for Voice and Speech*, 2000.

[2] John Laver. *The Phonetic Description of Voice Quality*. Cambridge University Press, Cambridge, 1980.

[3] John Kane and Christer Gobl. Wavelet maxima dispersion for breathy to tense voice discrimination. *IEEE Trans. Audio, Speech & Lang. Process.*, 21(6):1170–1179, 2013.

[4] Christer Gobl and Ailbhe Ní Chasaide. The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, 40(1-2):189–212, 2003.

[5] Erkki Vilkman, Eija-Riitta Lauri, Paavo Alku, Eeva Sala, and Marketta Sihvo. Loading changes in time-based parameters of glottal flow waveforms in different ergonomic conditions. *Folia Phoniatrica et Logopaedica*, 49(5):247–263, 1997.

[6] Erkki Vilkman. Voice problems at work: a challenge for occupational safety and health arrangement. *Folia Phoniatrica et Logopaedica*, 52(1-3):120–125, 2000.

[7] Eva Székely, John Kane, Stefan Scherer, Christer Gobl, and Julie Carson-Berndsen. Detecting a targeted voice style in an audiobook using voice quality features. In *Proc. ICASSP*, pages 4593–4596, March 2012.

[8] Donald G Childers and Chih K Lee. Vocal quality factors: Analysis, synthesis, and perception. *The Journal of the Acoustical Society of America*, 90(5):2394–2410, 1991.

[9] Axel Roebel, Stefan Huber, Xavier Rodet, and Gilles Degottex. Analysis and modification of excitation source characteristics for singing voice synthesis. In *Proc. ICASSP*, pages 5381–5384, 2012.

[10] Jaime Lorenzo-Trueba, Roberto Barra-Chicote, Tuomo Raitio, Nicolas Obin, Paavo Alku, Junichi Yamagishi, and Juan M Montero. Towards glottal source controllability in expressive speech synthesis. In *Proc. INTERSPEECH*, pages 1–1, 2012.

[11] Robert E. Hillman, Eva B. Holmberg, Joseph S. Perkell, Michael Walsh, and Charles Vaughan. Objective assessment of vocal hyperfunction. *Journal of Speech, Language, and Hearing Research*, 32(2):373–392, 1989.

[12] Daryush D Mehta, Jarrad H Van Stan, Matías Zañartu, Marzyeh Ghassemi, John V Guttag, Víctor M Espinoza, Juan P Cortés, Harold A Cheyne, and Robert E Hillman. Using ambulatory voice monitoring to investigate common voice disorders: Research update. *Frontiers in Bioengineering and Biotechnology*, 3:155, 2015.

[13] Nelson Roy, Julie Barkmeier-Kraemer, Tanya Eadie, M. Preeti Sivasankar, Daryush Mehta, Diane Paul, and Robert Hillman. Evidence-based clinical voice assessment: A systematic review. *American Journal of Speech-Language Pathology*, 22(2):212–226, 2013.

[14] Matthew Gordon and Peter Ladefoged. Phonation types: a cross-linguistic overview. *Journal of Phonetics*, 29(4):383–406, 2001.

[15] Mary Pietrowicz, Mark Hasegawa-Johnson, and Karrie G Karahalios. Acoustic correlates for perceived effort levels in male and female acted voices. *The Journal of the Acoustical Society of America*, 142(2):792–811, 2017.
[16] Matti Airas and Paavo Alku. Comparison of multiple voice source parameters in different phonation types. In *Proc. INTERSPEECH*, pages 1410–1413, 2007.
[17] Nick Campbell and Parham Mokhtari. Voice quality: the 4th prosodic dimension. In *Proc. ICPhS*, pages 2417–2420, 2003.
[18] Ioulia Grichkovtsova, Michel Morel, and Anne Lacheret. The role of voice quality and prosodic contour in affective speech perception. *Speech Communication*, 54(3):414–429, 2012.
[19] Soo Jin Park, Amber Afshan, Zhi Ming Chua, and Abeer Alwan. Using voice quality supervectors for affect identification. In *Proc. INTERSPEECH*, pages 157–161, 2018.
[20] Amber Afshan, Jinxi Guo, Soo Jin Park, Vijay Ravi, Jonathan Flint, and Abeer Alwan. Effectiveness of voice quality features in detecting depression. In *Proc. INTERSPEECH*, pages 1676–1680, 2018.
[21] Peter Birkholz, Lucia Martin, Klaus Willmes, Bernd J Kröger, and Christiane Neuschaefer-Rube. The contribution of phonation type to the perception of vocal emotions in German: an articulatory synthesis study. *The Journal of the Acoustical Society of America*, 137(3):1503–1512, 2015.
[22] Mika Ito. Politeness and voice quality-the alternative method to measure aspiration noise. In *Proc. Speech Prosody*, 2004.
[23] Irena Yanushevskaya, Christer Gobl, and Ailbhe Ní Chasaide. Voice quality and f0 cues for affect expression: implications for synthesis. In *Ninth European Conference on Speech Communication and Technology*, 2005.
[24] Peter Ladefoged, Ian Maddieson, and Michel Jackson. *Investigating Phonation Types in Different Languages*. Vocal Physiology: Voice Production, Mechanisms and Functions, New York: Raven Press, 1988.
[25] Jianjing Kuang and Patricia Keating. Vocal fold vibratory patterns in tense versus lax phonation contrasts. *The Journal of the Acoustical Society of America*, 136(5):2784–2797, 2014.
[26] Christina M Esposito. The effects of linguistic experience on the perception of phonation. *Journal of Phonetics*, 38(2):306–316, 2010.
[27] Sameer ud Dowla Khan. The phonetics of contrastive phonation in Gujarati. *Journal of Phonetics*, 40(6):780–795, 2012.
[28] Paavo Alku. 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, 2011.
[29] Paavo Alku, Juha Vintturi, and Erkki Vilkman. Measuring the effect of fundamental frequency raising as a strategy for increasing vocal intensity in soft, normal and loud phonation. *Speech Communication*, 38(3-4):321–334, 2002.
[30] Dhananjaya Gowda and Mikko Kurimo. Analysis of breathy, modal and pressed phonation based on low frequency spectral density. In *Proc. INTERSPEECH*, pages 3206–3210, 2013.
[31] James Hillenbrand, Ronald A Cleveland, and Robert L Erickson. Acoustic correlates of breathy vocal quality. *Journal of Speech, Language, and Hearing Research*, 37(4):769–778, 1994.
[32] Michal Borsky, Daryush D Mehta, Jarrad H Van Stan, and Jon Gudnason. Modal and nonmodal voice quality classification using acoustic and electroglottographic features. *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, 25(12):2281–2291, 2017.
[33] Thomas Drugman, Paavo Alku, Abeer Alwan, and Bayya Yegnanarayana. Glottal source processing: From analysis to applications. *Computer Speech and Language*, 28(5):1117–1138, 2014.
[34] Paavo Alku, Helmer Strik, and Erkki Vilkman. Parabolic spectral parameter - A new method for quantification of the glottal flow. *Speech Communication*, 22(1):67–79, 1997.
[35] Marc Swerts and Raymond N. J. Veldhuis. The effect of speech melody on voice quality. *Speech Communication*, 33(4):297–303, 2001.
[36] Marc Garellek, Robin Samlan, Bruce R. Gerratt, and Jody Kreiman. Modeling the voice source in terms of spectral slopes. *The Journal of the Acoustical Society of America*, 139(3):1404–1410, 2016.
[37] Jody Kreiman, Yen-Liang Shue, Gang Chen, Markus Iseli, Bruce R Gerratt, Juergen Neubauer, and Abeer Alwan. Variability in the relationships among voice quality, harmonic amplitudes, open quotient, and glottal area waveform shape in sustained phonation. *The Journal of the Acoustical Society of America*, 132:2625–2632, 2012.

[38] Dennis H Klatt and Laura C Klatt. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America*, 87(2):820–857, 1990.

[39] Polina Proutskova, Christophe Rhodes, Tim Crawford, and Geraint Wiggins. Breathy, resonant, pressed–automatic detection of phonation mode from audio recordings of singing. *Journal of New Music Research*, 42(2):171–186, 2013.

[40] Johan Sundberg. The perception of singing. In *The Psychology of Music (Second Edition)*, pages 171–214. Elsevier, 1999.

[41] Johan Sundberg. *The Science of the Singing Voice*. Illinois University Press, 1987.

[42] Johan Sundberg. The acoustics of the singing voice. *Scientific American*, 236:82–91, 1977.

[43] Elizabeth U Grillo and Katherine Verdolini. Evidence for distinguishing pressed, normal, resonant, and breathy voice qualities by laryngeal resistance and vocal efficiency in vocally trained subjects. *Journal of Voice*, 22(5):546–552, 2008.

[44] Johan Sundberg. Vocal fold vibration patterns and modes of phonation. *Folia Phoniatrica et Logopaedica*, 47(4):218–228, 1995.

[45] Ninni Elliot, Johan Sundberg, and Patricia Gramming. Physiological aspects of a vocal exercise. *Journal of Voice*, 11(2):171–177, 1997.

[46] Jackie L Gartner-Schmidt, Douglas F Roth, Thomas G Zullo, and Clark A Rosen. Quantifying component parts of indirect and direct voice therapy related to different voice disorders. *Journal of Voice*, 27(2):210–216, 2013.

[47] Jean-Luc Rouas and Leonidas Ioannidis. Automatic classification of phonation modes in singing voice: Towards singing style characterisation and application to ethnomusicological recordings. In *Proc. INTERSPEECH*, pages 150 – 154, 2016.

[48] Ixone Arroabarren and Alfonso Carlosena. Inverse filtering in singing voice: a critical analysis. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1422–1431, July 2006.

[49] Ixone Arroabarren and Alfonso Carlosena. Vibrato in singing voice: the link between source-filter and sinusoidal models. *EURASIP Journal on Applied Signal Processing*, 2004:1007–1020, 2004.

[50] Moa Millgrd, Tobias Fors, and Johan Sundberg. Flow glottogram characteristics and perceived degree of phonatory pressedness. *Journal of Voice*, 30(3):287–292, 2016.

[51] Daniel Stoller and Simon Dixon. Analysis and classification of phonation modes in singing. In *Proc. International Society for Music Information Retrieval*, 2016.

[52] Jody Kreiman, Soo Jin Park, Patricia A. Keating, and Abeer Alwan. The relationship between acoustic and perceived intraspeaker variability in voice quality. In *Proc. INTERSPEECH*, pages 2357–2360, 2015.

[53] Sudarsana Reddy Kadiri and Bayya Yegnanarayana. Breathy to tense voice discrimination using zero-time windowing cepstral coefficients (ztwccs). In *Proc. INTERSPEECH*, pages 232–236, 2018.

[54] Sudarsana Reddy Kadiri and Bayya Yegnanarayana. Analysis and detection of phonation modes in singing voice using excitation source features and single frequency filtering cepstral coefficients (sffcc). In *Proc. INTERSPEECH*, pages 441–445, 2018.

[55] K. Sri Rama Murty and Bayya Yegnanarayana. Epoch extraction from speech signals. *IEEE Trans. Audio, Speech, and Lang. Process.*, 16(8):1602–1613, Nov. 2008.

[56] Bayya Yegnanarayana and Dhananjaya Gowda. Spectro-temporal analysis of speech signals using zero-time windowing and group delay function. *Speech Communication*, 55(6):782–795, 2013.

[57] G. Aneeja and Bayya Yegnanarayana. Single frequency filtering approach for discriminating speech and non-speech. *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, 23(4):705–717, Apr. 2015.

[58] P. Gangamohan, Sudarsana Reddy Kadiri, and Bayya Yegnanarayana. Analysis of emotional speech at subsegmental level. In *Proc. INTERSPEECH*, pages 1916–1920, Aug. 2013.

[59] Sudarsana Reddy Kadiri, P. Gangamohan, Suryakanth V Gangashetty, and Bayya Yegnanarayana. Analysis of excitation source features of speech for emotion recognition. In *Proc. INTERSPEECH*, pages 1324–1328, 2015.

[60] Paavo Alku, Tom Bäckström, and Erkki Vilkman. Normalized amplitude quotient for parametrization of the glottal flow. *The Journal of the Acoustical Society of America*, 112(2):701–710, Feb. 2002.

[61] Guruprasad Seshadri and Bayya Yegnanarayana. Perceived loudness of speech based on the characteristics of glottal excitation source. *The Journal of the Acoustical Society of America*, 126:2061–2071, 2009.

[62] Ravi Shankar Prasad and Bayya Yegnanarayana. Determination of glottal open regions by exploiting changes in

the vocal tract system characteristics. *The Journal of the Acoustical Society of America*, 140(1):666–677, 2016.
[63] Sudarsana Reddy Kadiri, RaviShankar Prasad, and B. Yegnanarayana. Detection of glottal closure instant and glottal open region from speech signals using spectral flatness measure. *Speech Communication*, 116:30–43, 2020.
[64] Kenneth N Stevens. *Acoustic Phonetics*, volume 30. MIT press, Cambridge, MA, 2000.
[65] Anna Barney, Antonio De Stefano, and Nathalie Henrich. The effect of glottal opening on the acoustic response of the vocal tract. *Acta Acustica united with Acustica*, 93(6):1046–1056, 2007.
[66] Sudarsana Reddy Kadiri and Bayya Yegnanarayana. Epoch extraction from emotional speech using single frequency filtering approach. *Speech Communication*, 86:52 – 63, 2017.
[67] John D Markel and Augustine H. Gray Jr. *Linear Prediction of Speech*, volume 12. Springer Science & Business Media, 2013.
[68] K. N. R. K. Raju Alluri, Sivanand Achanta, Sudarsana Reddy Kadiri, Suryakanth V. Gangashetty, and Anil Kumar Vuppala. SFF anti-spoofer: IIIT-H submission for automatic speaker verification spoofing and countermeasures challenge 2017. In *Proc. INTERSPEECH*, pages 107–111, 2017.
[69] John Kane and Christer Gobl. Identifying regions of non-modal phonation using features of the wavelet transform. In *Proc. INTERSPEECH*, pages 177–180, 2011.
[70] Polina Proutskova. Singing phonation database.
[71] Paavo Alku. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech Communication*, 11(2-3):109–118, June 1992.
[72] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. Covarep – a collaborative voice analysis repository for speech technologies. In *Proc. ICASSP*, pages 960–964, 2014.
[73] Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.