Interspeech 2018
2-6 September 2018, Hyderabad

Breathy to Tense Voice Discrimination using Zero-Time Windowing Cepstral Coefficients (ZTWCCs)

Sudarsana Reddy Kadiri and B. Yegnanarayana

Speech Processing Laboratory,
International Institute of Information Technology, Hyderabad, India
sudarsanareddy.kadiri@research.iiit.ac.in, yegna@iiit.ac.in

Abstract

In this paper, we consider breathy to tense voices, which are often considered to be opposite ends of a voice quality continuum. Along with these, other aspects of a speaker's voice play an important role to convey the information to the listener such as mood, attitude and emotional state. The glottal pulse characteristics in different phonation types vary due to the tension of laryngeal muscles together with the respiratory effort. In the present study, we are deriving the features that can capture effects of excitation on the vocal tract system through a signal processing method, called as zero-time windowing (ZTW) method. The ZTW method gives the instantaneous spectrum which captures the changes in the speech production mechanism, providing higher spectral resolution. The cepstral coefficients derived from ZTW method are used for the classification of phonation types. Along with zero-time windowing cepstral coefficients (ZTWCCs), we use the excitation source features derived from zero frequency filtering (ZFF) method. The excitation features used are: strength of excitation, energy of excitation, loudness measure and ZFF signal energy. Classification experiments using ZTWCC and excitation features reveal a significant improvement in the detection of phonation type compared to the existing voice quality features and MFCC features.
Index Terms: Speech analysis, Excitation source, Phonation type.

1. Introduction

Voice quality or phonation type is considered as timbre or auditory coloring of a speaker's voice [1]. In the current study, we consider breathy to tense voice, which are often considered to be opposite ends of a voice quality continuum. Along with these voice qualities, other aspects of a speaker voice such as rhythm, timbre, intonation, intensity etc., play an important role to convey the information such as mood, attitude and emotional state [2, 3]. For example, studies in [4] shown that breathiness has been associated with expressing politeness and also intimacy and familiarity. On the other hand, tense voice has often been associated in more active (arousal) emotional states such as anger and happy emotions [5, 6]. Analysis, representation and detection of different voice qualities is desirable for various applications. It is helpful for tagging the voice qualities in highly expressive speech corpora [7]. The characterization of voice quality is required for speech synthesis and voice quality modification systems [8–10]. Detection of the type of voice quality may improve the performance of various speech processing applications like cognitive load of a speaker, speaker recognition, speech recognition, emotion recognition [11–19] etc. Note that, voice quality also has a phonological contrastive function in many languages [20, 21]

According to [1], voice qualities are compared with respect to modal/neutral voice. In modal voice, the laryngeal tension settings are in low and moderate range. The vocal folds vibrations are mostly periodic with a minimum irregularity in a sequence of glottal cycles with a proper glottal closure (no audible frication noise). Breathy voice typically involves weaker levels of laryngeal tension, partial closure of the glottis and often a glottal chink. These settings lead to generation of some amount of turbulence/aspiration noise at the vocal folds. As a result of aspiration noise, lower frequency harmonics are effected compared to modal voice. On the other hand, tense voice involves increase in longitudinal and adductive tension in terms of laryngeal settings. However, this tension does not bring the irregularities in vocal folds vibration, such as the case in harsh voice. The abrupt/sharpness of glottal closure characteristics of tense voice are reflected as stronger high frequency harmonics.

The characteristics of the glottal source varies in different phonation types due to the tension of the laryngeal muscles together with the respiratory effort. Hence, the glottal source waveform varies from a smooth symmetric form (typical case for soft voices) to asymmetric waveform with sharp edges such as occurs in the production of loud/pressed voices [22, 23]. This kind of time-domain variation is reflected as the decay of the spectral envelope of glottal source in the frequency-domain [24, 25]. Apart from these, there are also other parameterization methods, where the estimated glottal source waveform is matched with the glottal flow models such as LF model to obtain the model parameters [6, 26].

In [22,27], various types of glottal source parameters (time-based, amplitude-based and frequency-domain parameters) are analyzed for discriminating breathy, modal and tense voices. Frequency-domain parameters such as H1-H2 [25], harmonic richness factor (HRF) [8] and parabolic spectral parameter (PSP) [28] are derived by fitting a parabola to the lower frequencies of the glottal source spectrum. Time domain parameters such as closing quotient, quasi-open quotient (QOQ), open quotient and the speed quotient and amplitude-based parameters such as normalized amplitude quotient (NAQ) are derived from the glottal flow and its derivative waveforms [11, 22, 29, 30]. Also, studies [25, 31] measured the amount of aspiration noise present in the signal for the detection of breathy voice, based on the observation that third formant region to be considerably noisier in breathy compared to modal voice. In [6, 26], parameters derived after fitting the glottal flow model (such as LF model) with the estimated glottal source were analyzed for various voice qualities.

From the studies [24,32], it was observed that NAQ and H1-H2 are the best features for the identification of phonation types. While NAQ feature gives a measure of the skewness of the glottal pulse, H1-H2 feature captures its acoustic manifestation in the frequency-domain. It was observed that, the effectiveness of

10.21437/Interspeech.2018-2498

the glottal source parameters reduce for high pitch and expressive voices [11, 30]. To overcome this effect, recently attempts were made to extract features directly from the speech signal. In [32], to capture the sharp changes in glottal closure characteristics, a parameter called maximum dispersion quotient (MDQ) is proposed, which is derived from LP residual signal. In [24], the production characteristics such as breathy voice having higher open quotient and pressed voice having a least open quotient were captured using spectral-domain parameter called low frequency spectral density (LFSD). The effect on subglottal system in the spectrum is higher for breathy voice (owing to higher open quotient) compared to pressed/tense voice. This results in increase of low frequency spectral energy for breathy voice, typically around the region of the glottal formant (lower than first formant). From the studies [24], it was observed that LFSD and MDQ are close to NAQ, and HNR seems to provide less discrimination among three phonation types. However, HNR was shown to provide good performance in the discrimination of modal and breathy voices compared to modal and pressed voices. Also, it was observed that H1-H2 performs poor for female speakers and it is as good as NAQ for male speakers. This may be due to the overlap of second harmonic with the first formant for female voice. In studies [24, 32] authors used a set of voice quality features such as NAQ, QOQ, H1-H2, PSP and MDQ for the classification of phonation. From the analysis, it was observed that no single feature performs consistently better for all the speakers in the discrimination of phonation types. Hence, there is a need for exploring alternative features for the analysis and classification of voices such as phonations types from the speech signal. In this paper, we explore the features that reflect the effect of excitation on vocal tract system through cepstral coefficients derived from ZTW method [33] and excitation features derived from ZFF method [34].

The organization of the paper is as follows. Section 2 describes the signal processing methods used for feature extraction. Section 3 describes the analysis of excitation source features. In Section 4, we discuss the experimental protocol which includes the databases and features used for comparison. Details of classification experiments and discussion on results are presented in Section 5. Finally, Section 6 gives a summary of the study.

2. Signal Processing Methods for Feature Extraction

In this section, we describe the features that reflect the effect of excitation on the vocal tract system, derived from the zero time windowing (ZTW) method [33]. We also use excitation source features which are derived from ZFF method [34]. It is to be noted that, either of these two signal processing methods do not assume source-filter model of speech production mechanism.

2.1. ZTW Method and Extraction of ZTWCC

The objective of the method is to derive the instantaneous spectral characteristics, so that the time varying characteristics of the speech production mechanism can be captured. In this method, the speech signal is windowed with a heavily decaying window (which provides high emphasis to the samples near the starting sampling instant, which is referred as zero time) gives high temporal resolution, whereas the group delay provides good resolution of the spectral characteristics. Hence, the method provides higher temporal resolution, simultaneously maintaining good spectral resolution. The ZTW spectrum was shown to capture the excitation variations such as glottal opening, open phase and also time varying system characteristics such as vocal tract resonances effectively [33, 35].

The steps involved in extracting the instantaneous spectral characteristics using the ZTW method [33] are as follows.

1. Consider a L $msec$ speech segment $s[n]$ (number of samples, $M = L * f_s/1000$) at each instant. The segment is appended with sufficient number of zeros before computing N-point DFT for observing spectral characteristics. In this study $N = 2048$ is used.
2. Multiply $s[n]$ segment with a window $w_1^2[n]w_2[n]$, where

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

$$w_2[n] = 4\, cos^2(\pi n/2M), \quad n = 0, \ldots, M-1.$$

   Multiplying the signal with the heavily decaying window $w_1^2[n]$ is called zero time windowing, which emphasizes the values near the beginning of the window. The window $w_1^2[n]$ gives approximately four times integration in the frequency domain.
3. The numerator of group delay (NGD) function ($g[k]$) of the windowed signal (i.e., of $x[n] = w_1^2[n]w_2[n]s[n]$) is computed to estimate the spectrum and is given by

$$g[k] = X_R[k]Y_R[k] + X_I[k]Y_I[k], \quad k = 0, \ldots, N-1.$$

   where $X_R[k]$ and $X_I[k]$ are the real and imaginary parts of the $X[k]$, respectively, where $X[k]$ is the N-point DFT of $x[n]$. $Y_R[k]$ and $Y_I[k]$ are the real and imaginary parts of the $Y[k]$, respectively, where $Y[k]$ is the N-point DFT of $y[n]$ ($y[n] = nx[n]$).
4. The resulting NGD function is double differenced ($g''[k]$) to highlight the spectral features such as resonances/formants of the vocal tract system.
5. The low amplitude peaks in the double differenced NGD plots are highlighted by computing its Hilbert envelope. The resulting spectrum is called the HNGD spectrum, and it is denoted as $S[n, k]$ in this study.

Figure 1 gives an illustration of HNGD spectrograms for breathy, modal and pressed phonations. It can be clearly seen that there exists significant spectral variations due to excitation effect on the system. In order to capture these variations, we propose to derive the zero-time windowing cepstral coefficients (ZTWCCs).

2.1.1. ZTWCC Extraction

Cepstrum $c[n, k]$ is computed from ZTW/HNGD spectrum $S[n, k]$, and is given by

$$c[n, k] = IFFT(log(S[n, k])).$$

From $c[n, k]$, first 13 cepstral coefficients are considered and they are named as ZTWCCs. The ZTWCCs can be obtained at each time instant. In this study, instead of computing at each instant, we computed the ZTWCCs at glottal closure instant (GCI) locations. From static (S) coefficients, delta (V) and double-delta (A) coefficients are also computed, which makes total of 39 dimension. The schematic block diagram of ZTWCCs extraction is shown in Fig. 2.

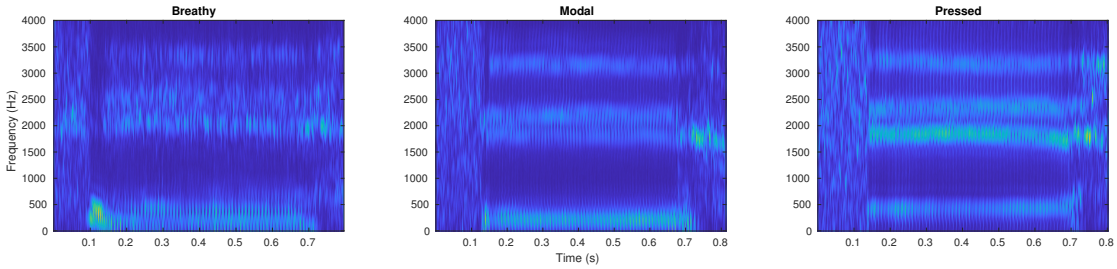

Figure 1: *An illustration of HNGD spectrograms for breathy, modal and pressed phonations.*

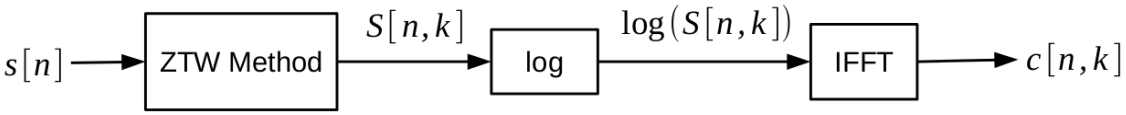

Figure 2: *Block diagram of zero-time windowing cepstral coefficients (ZTWCCs) extraction.*

2.2. Zero frequency filtering (ZFF) method

The ZFF [34] method gives the robust estimates of glottal closure instants (GCIs). The method is also useful for deriving features such as instantaneous fundamental frequency and strength of impulse-like excitation. The method relies on observation that impulsive nature of the excitation is reflected across all frequencies including zero frequency (0 Hz). Hence, the GCI locations are detected by confining the analysis around 0 Hz. In this method, the pre-emphasized speech signal $x[n]$ is passed through a cascade of two ideal zero-frequency resonators. The resulting signal grows/decays approximately as a polynomial function of time. The trend is removed by subtracting the local mean at each sample. The resulting signal is the zero-frequency filtered (ZFF) signal. The locations of negative-to-positive zero crossings (NPZCs) correspond to GCIs.

Features of excitation source: From the ZFF method, we derive the excitation features such as strength of excitation (SoE), energy of excitation (EoE), perceived loudness and ZFF signal energy. The SoE is computed as the slope of the ZFF signal around each NPZCs which is proportional to the rate of closure of the vocal folds [36, 37]. The energy of excitation (EoE) parameter is computed from the samples of the Hilbert envelope of LP residual over 2 ms region around each GCI. This gives a measure of the vocal effort [37]. A 10^{th} order LP analysis is used for each frame of 16 ms and a frame shift of 2 ms. Loudness (perceived loudness) measure captures the abruptness of glottal closure [38]. It is defined as the ratio of standard deviation and mean of the samples of the Hilbert envelope of LP residual signal around GCI. The other excitation parameter is the energy of the ZFF signal, which is computed for a frame size of 10 ms with a sample shift. The energy of the ZFF signal at GCI is considered in this study.

3. Feature analysis

The distributions of the proposed excitation source features mentioned above are given Fig. 3. It can be seen that SoE values are high for breathy and low for pressed/tense voice. EoE values are high for pressed voice and low for breathy voice. This is because the vocal effort is more in the case of pressed phonation than breathy voice. The perceived loudness measure values for breathy voice are lower than modal and pressed phonations. The ZFF signal energy comes out to be lower for pressed followed by modal and breathy. In summary, it can be observed that there exists a significant discrimination among the feature values for all the phonation types.

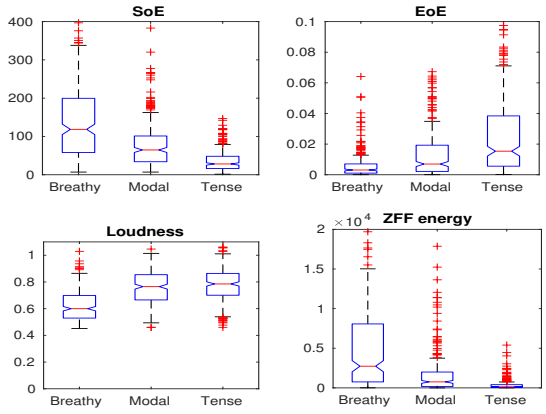

Figure 3: *Distribution of features for breathy, modal and pressed phonation types using box plots. The central mark indicates the median, and the bottom and top edges of the box indicate the 25^{th} and 75^{th} percentiles, respectively. The whiskers extend to the most extreme data points not considered outliers, and the outliers are plotted individually using the $'+'$ symbol.*

4. Experimental protocol

This section describes the phonation database used in the study and the features (voice quality features and MFCCs) used for comparison with proposed features (ZTWCCs and excitation features derived from ZFF method).

4.1. Database used

The phonation database used in this study consists of 8 different Finnish vowels uttered in three phonation types (breathy, modal and pressed) by 6 female and 5 male speakers (aged between 18 and 48 years). Each vowel is uttered three times making it a total of 792 (3*3*8*11) isolated vowels. The database was

originally recorded at a sampling frequency of 44.1 kHz in an anechoic chamber and later it is downsampled to 16 kHz. More details of the database can be found in [22, 39].

4.2. Features for comparison

Voice quality features and MFCCs are considered for comparison. The selection of these features is based on the findings in [22, 24, 32], which is shown to be the most suitable measurements for discrimination of breathy to tense voice. The voice quality feature set consists of NAQ, QOQ, H1-H2, PSP and MDQ. Among these, first four features are derived from Iterative and Adaptive Inverse Filtering (IAIF) method [40]. A brief description of the features is given below.

Normalized Amplitude Quotient (NAQ) [29]: It is the ratio of the AC-amplitude of the glottal flow (noted f_{AC}) and the negative peak amplitude d_{min} of the glottal flow derivative, normalized with the pitch period.
H1-H2: It is the difference between the amplitudes of the first & second harmonics of the glottal flow spectrum [22].
Quasi-open quotient (QOQ) [11, 22]: It is calculated by detecting the peak in the glottal flow and finding the time points previous to and following that descend below 50% of the peak amplitude. The duration between the time locations gives as a quasi-open phase and divided by the local pitch period gives QOQ. This parameter is closely related to the open quotient.
Maximum dispersion quotient (MDQ) [32]: This parameter measures the dispersion in the LP residual around the GCI.
Parabolic spectral parameter (PSP) [28]: PSP is a frequency domain feature developed for the quantification of the glottal flow waveform.
Mel-frequency cepstral coefficients (MFCCs): In this study, 13 mel-frequency cepstral coefficients are measured using 25 ms Hamming windowed frames, with a 5 ms shift. From static (S) coefficients, delta (V) and double-delta (A) coefficients were computed, which makes total of 39 dimension.

5. Classification Experiments and Discussion on Results

The classification experiments are carried out using support vector machines (SVMs) utilizing a radial basis function (RBF) kernel [41]. Experiments are conducted using 10-fold cross-validation, where the dataset is randomly partitioned into 10 equal sets. One fold is held out to be used for testing with the remaining dataset for training. This process is repeated for each of the 10-folds (classification accuracies are saved in each fold). The experiments are carried out for 6 different feature vectors:

- VQ = [NAQ,QOQ, H1-H2, PSP and MDQ]
- MFCC
- Excitation = [SoE, EoE, Loudness, ZFF energy]
- ZTWCC
- Excitation+ZTWCC
- VQ+MFCC+Excitation+ZTWCC

The classification results from the 10-fold cross-validation experiments are shown in terms of mean and standard deviation of the classification accuracies in Table 1. From table, it can be seen that including all parameters (i.e., VQ features, MFCCs, excitation features, and ZTWCC) gives the highest average classification accuracy (75.31%). It can be also be observed that phonation classification accuracy with ZTWCC gives the highest performance compared to VQ features and MFCCs. Even though excitation features alone are not showing significant classification accuracy, by combining with ZTWCC gives the significant improvement in accuracy. This indicates excitation features and ZTWCC are providing complimentary information.

Table 1: *Mean and standard deviation of classification accuracy scores (in %) after 10-fold cross validation with different input feature vectors.*

Features	**Mean accuracy**[%]	**Standard deviation**[%]
VQ	64.21	4.97
MFCC	68.52	5.14
Excitation	61.26	5.84
ZTWCC	69.38	4.53
Excitation+ZTWCC	72.37	4.18
VQ+MFCC+Excitation+ZTWCC	**75.31**	4.11

Table 2: *Confusion matrix (in %) with 10-fold cross validation after combining all features (i.e., VQ+MFCC+Excitation+ZTWCC).*

	Breathy [%]	Modal [%]	Tense [%]
Breathy	84.61	13.77	1.62
Modal	13.21	66.66	20.13
Tense	4.06	22.14	73.80

The confusion matrix displayed in Table 2 shows that there is a significant accuracy for breathy voice and confusion between modal and tense voice when all features are used for classification. The trend in the accuracies among the phonation classes is in conformity with the studies [24, 32]. The classification accuracy can be further improved by including other voice quality parameters and by exploring features that can capture the effect of excitation on the system characteristics. Also, to ensure the quality of the intended phonation type and to resolve the potential ambiguity, there is a need for perceptual screening of the data.

6. Summary and conclusion

In this paper, we present new features, zero-time windowing cepstral coefficients (ZTWCC) for discriminating breathy to tense voice. Along with ZTWCC, we also derived excitation source features for the analysis and classification. A comprehensive evaluation reveals that the proposed features provides better differentiation of the phonation classes. The existing voice quality features (except MDQ) are calculated from the glottal source waveform estimated by inverse filtering. However, automatic glottal inverse filtering for continuous speech can be problematic for high pitched voices. Proposed ZTWCC and excitation source features do not assume source-filter model of speech production and do not use glottal inverse filtering. This suggests that, the proposed features are more suitable for automatic analysis of continuous speech and high pitched voices [42]. Results from the classification experiments clearly demonstrate that ZTWCC and excitation source features provide further information than that is present in existing voice quality parameters and MFCCs, for discriminating phonation types. Furthermore, experiments showed that on its own ZTWCC can be used to achieve a lower classification error than other individual features.

7. References

[1] J. Laver, *The Phonetic Description of Voice Quality*. Cambridge: Cambridge University Press, 1980.

[2] N. Campbell and P. Mokhtari, "Voice quality: the 4th prosodic dimension," in *15th ICPhS*, 2003, pp. 2417–2420.

[3] I. Grichkovtsova, M. Morel, and A. Lacheret, "The role of voice quality and prosodic contour in affective speech perception," *Speech Communication*, vol. 54, no. 3, pp. 414–429, 2012.

[4] M. Ito, "Politeness and voice quality-the alternative method to measure aspiration noise," in *Speech Prosody 2004, International Conference*, 2004.

[5] I. Yanushevskaya, C. Gobl, and A. N. Chasaide, "Voice quality and f0 cues for affect expression: implications for synthesis," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[6] C. Gobl and A. N. Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Communication*, vol. 40, no. 1-2, pp. 189–212, 2003.

[7] . Szkely, J. Kane, S. Scherer, C. Gobl, and J. Carson-Berndsen, "Detecting a targeted voice style in an audiobook using voice quality features," in *ICASSP*, March 2012, pp. 4593–4596.

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

[9] A. Roebel, S. Huber, X. Rodet, and G. Degottex, "Analysis and modification of excitation source characteristics for singing voice synthesis," in *ICASSP*, 2012, pp. 5381–5384.

[10] J. Lorenzo-Trueba, R. Barra-Chicote, T. Raitio, N. Obin, P. Alku, J. Yamagishi, and J. M. Montero, "Towards glottal source controllability in expressive speech synthesis," in *Interspeech*, 2012, pp. 1–1.

[11] T. Drugman, P. Alku, A. Alwan, and B. Yegnanarayana, "Glottal source processing: From analysis to applications," *Computer Speech & Language*, vol. 28, no. 5, pp. 1117 – 1138, 2014.

[12] M. Airas and P. Alku, "Emotions in vowel segments of continuous speech: Analysis of the glottal flow using the normalised amplitude quotient," *Phonetica*, vol. 63, no. 1, pp. 26–46, 2006.

[13] M. Tahon, G. Degottex, and L. Devillers, "Usual voice quality features and glottal features for emotional valence detection," in *Proc. International Conference on Speech Prosody*, Shanghai, China, May 2012.

[14] J. Sundberg, S. Patel, E. Björkner, and K. R. Scherer, "Interdependencies among voice source parameters in emotional speech," *T. Affective Computing*, vol. 2, no. 3, pp. 162–174, 2011.

[15] M. Lugger and B. Yang, "Cascaded emotion classification via psychological emotion dimensions using a large set of voice quality parameters," in *ICASSP*, 2008, pp. 4945–4948.

[16] T. F. Yap, J. Epps, E. Ambikairajah, and E. H. C. Choi, "Voice source features for cognitive load classification," in *ICASSP*, 2011, pp. 5700–5703.

[17] K. W. Godin, T. Hasan, and J. H. L. Hansen, "Glottal waveform analysis of physical task stress speech," in *INTERSPEECH*, 2012.

[18] M. B. A. K. S. K. H. J. C. R. E. Shriberg, M. Graciarena and F. Goodman, "Effects of vocal effort and speaking style on text-independent speaker verification," in *INTERSPEECH*, 2008, pp. 609–612.

[19] M. S. P. Zelinka and J. Schimmel, "Impact of vocal effort variability on automatic speech recognition," *Speech Communication*, vol. 54, no. 6, p. 732742, 2012.

[20] I. M. Ladefoged, Peter and M. Jackson, *Investigating Phonation Types in Different Languages*. New York: Raven Press, 1988.

[21] M. Gordon and P. Ladefoged, "Phonation types: a cross-linguistic overview," *Journal of Phonetics*, vol. 29, no. 4, pp. 383–406, 2001.

[22] M. Airas and P. Alku, "Comparison of multiple voice source parameters in different phonation types," in *INTERSPEECH*, 2007, pp. 1410–1413.

[23] P. Alku, J. Vintturi, and E. Vilkman, "Measuring the effect of fundamental frequency raising as a strategy for increasing vocal intensity in soft, normal and loud phonation," *Speech Communication*, vol. 38, no. 3-4, pp. 321–334, 2002.

[24] D. Gowda and M. Kurimo, "Analysis of breathy, modal and pressed phonation based on low frequency spectral density," in *INTERSPEECH*, 2013, pp. 3206–3210.

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

[26] M. Swerts and R. N. J. Veldhuis, "The effect of speech melody on voice quality," *Speech Communication*, vol. 33, no. 4, pp. 297–303, 2001.

[27] J. Kane and C. Gobl, "Evaluation of glottal closure instant detection in a range of voice qualities," *Speech Communication*, vol. 55, no. 2, pp. 295–314, 2013.

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

[29] P. Alku, T. Backstrom, and E. Vilkman, "Normalized amplitude quotient for parametrization of the glottal flow," *The Journal of the Acoustical Society of America*, vol. 112, no. 2, pp. 701–710, Feb. 2002.

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

[31] D. Klatt and L. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Amer.*, vol. 87, no. 2, pp. 820–857, 1990.

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

[33] B. Yegnanarayana and N. G. Dhananjaya, "Spectro-temporal analysis of speech signals using zero-time windowing and group delay function," *Speech Communication*, vol. 55, no. 6, pp. 782–795, 2013.

[34] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1602–1613, Nov. 2008.

[35] R. S. Prasad and B. Yegnanarayana, "Determination of glottal open regions by exploiting changes in the vocal tract system characteristics," *The Journal of the Acoustical Society of America*, vol. 140, no. 1, pp. 666–677, 2016.

[36] P. Gangamohan, S. R. Kadiri, and B. Yegnanarayana, "Analysis of emotional speech at subsegmental level," in *INTERSPEECH*, Aug. 2013, pp. 1916–1920.

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

[38] S. Guruprasad and B. Yegnanarayana, "Performance of an event-based instantaneous fundamental frequency estimator for distant speech signals," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 7, pp. 1853–1864, Sept 2011.

[39] J. Kane and C. Gobl, "Identifying regions of non-modal phonation using features of the wavelet transform," in *INTERSPEECH*, 2011, pp. 177–180.

[40] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Communication*, vol. 11, no. 2-3, pp. 109–118, June 1992.

[41] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," vol. 2, July 2007.

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