INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Mel-frequency Cepstral Coefficients of Voice Source Waveforms for Classification of Phonation Types in Speech

Sudarsana Reddy Kadiri, Paavo Alku

Department of Signal Processing and Acoustics, Aalto University, Finland
{sudarsana.kadiri,paavo.alku}@aalto.fi

Abstract

Voice source characteristics in different phonation types vary due to the tension of laryngeal muscles along with the respiratory effort. This study investigates the use of mel-frequency cepstral coefficients (MFCCs) derived from voice source waveforms for classification of phonation types in speech. The cepstral coefficients are computed using two source waveforms: (1) glottal flow waveforms estimated by the quasi-closed phase (QCP) glottal inverse filtering method and (2) approximate voice source waveforms obtained using the zero frequency filtering (ZFF) method. QCP estimates voice source waveforms based on the source-filter decomposition while ZFF yields source waveforms without explicitly computing the source-filter decomposition. Experiments using MFCCs computed from the two source waveforms show improved accuracy in classification of phonation types compared to the existing voice source features and conventional MFCC features. Further, it is observed that the proposed features have complimentary information to the existing features.

Index Terms: Speech analysis, Voice source, Phonation type, Voice quality, Glottal inverse filtering, Zero frequency filtering.

1. Introduction

Voice quality, defined as auditory coloring of a person's voice [1], is affected by the phonation type of the vocal folds in speech production. Breathy and tense voice are considered to be the opposite ends of the voice quality continuum. Along with voice quality, also other aspects such as rhythm, intonation and intensity convey expressive characteristics (e.g. mood, affect and emotional state of the speaker) in speech signals [2–4]. In [5], it was reported that breathiness is associated with expression of politeness, intimacy and familiarity. Tense voice, however, has been associated with emotions of high arousal such as anger and happiness [6, 7]. Analysis and detection of different phonation types is desirable for various applications in speech and voice research such as tagging voice qualities in expressive speech corpora, in speech synthesis and in voice modification systems [8–11]. Phonation type classification can also improve the performance of various speech processing applications like assessment of speech pathology and cognitive load of the speaker, speech recognition, speaker recognition and emotion recognition [12–20].

According to Laver [1], variations in laryngeal activity and changes in the laryngeal settings give rise to different phonation types. In comparing different phonation types, modal phonation is typically used as the reference to which other types are compared [1]. Modal phonation is the most efficient type of phonation produced using vocal fold vibration which is of moderate adductive and longitudinal tension and of moderate medial compression. In modal phonation, the vocal folds vibrate quasi-periodically with minimal frication and complete glottal closure. In breathy phonation, the vocal folds vibrate in a more inefficient manner and the vibration is accompanied by audible friction. The reduced muscular tension allows a constant turbulent air to pass through the glottis. Tense voice is produced using increased laryngeal tension compared to modal voice. Tense phonation is also characterized by sharper closure of the glottis.

The above mentioned differences in functioning of the vocal folds affect the time domain waveform of the acoustical excitation generated by the vocal folds, the glottal flow. Hence, the glottal flow waveform varies from a smooth, almost symmetric form in breathy phonation to a more asymmetric waveform with sharp edges at glottal closure in tense phonation [21, 22]. This type of time domain variation is reflected in the frequency domain as variation in the spectral tilt of the glottal flow: breathy phonation shows a steep spectral tilt whereas tense phonation shows a smaller spectral tilt [23, 24]. As a result of increased friction, lower harmonics are strong in breathy phonation. The increased sharpness of glottal closure characteristics in tense phonation results in more prominent upper harmonics.

In order to capture variations in the glottal flow waveform, different parameterization methods have been developed. These methods can be divided into time domain and frequency domain parameters, and the former category can further be divided to time-based and amplitude-based measures [25]. Examples of time-based parameters are the open quotient (OQ), the quasi-open quotient (QOQ), the speed quotient (SQ) and the closing quotient (CQ) [25]. The normalized amplitude quotient (NAQ) is an example of amplitude-based measures [26]. The level difference between the first two harmonics (H1-H2) [27], the harmonic richness factor (HRF) [9] and the parabolic spectral parameter (PSP) [28] are examples of frequency domain measures. It has been reported that NAQ and H1-H2 are effective features for the identification of the phonation type [23,29]. NAQ captures the skewness of the glottal pulse and H1-H2 captures the corresponding manifestation in the spectral tilt of the glottal pulse. Differently from the above studies, the estimated glottal flow waveforms were matched with the Liljencrants-Fant (LF) model in [7, 30] to obtain parameters for discriminating voice qualities. Studies in [24,31] measured the amount of friction noise for the detection of breathy voices based on the observation that the third formant region is considerably noisier in breathy phonation compared to modal phonation. In [24, 32], cepstral peak prominence (CPP) was shown to correlate well with breathiness.

To capture abrupt glottal closure characteristics in speech production, the peak slope and maximum dispersion quotient (MDQ) were derived from wavelet transform in [29, 33]. The production characteristics between breathy and pressed phonation were captured using the low-frequency spectral density (LFSD) in [23]. It was observed in [23] that discrimination capabilities of LFSD and MDQ are close to that of NAQ. Harmonic-to-noise ratio (HNR) was found to provide poor dis-

http://dx.doi.org/10.21437/Interspeech.2019-2863

crimination among phonation types in [23]. However, HNR was shown to discriminate modal and breathy voices better compared to modal and pressed voices. NAQ, QOQ, H1-H2, PSP and MDQ were used for the classification of phonation types from speech in [23, 29]. It was observed that no single feature performed consistently better than the other features for all the speakers. Hence, alternative features for the analysis and classification of phonation types are needed. Mel-frequency cepstral coefficients (MFCCs) derived from speech signals were investigated in [29, 34] for classification of phonation types in speech. Similarly the authors of [35] proposed cepstral features derived from high-resolution spectrum obtained by the zero-time windowing (ZTW) method. In this paper, we propose to derive MFCCs from voice source waveforms for classification of phonation types, as the voice source waveform contains significant information about phonation types.

The paper is organized as follows. Section 2 describes the signal processing methods used for deriving voice source waveforms and the extraction of cepstral coefficients from the computed voice source waveforms. Section 3 describes the experimental protocol including the database, features and classifier. Results and discussion on classification experiments are presented in Section 4. A summary of the paper is given in Section 5.

2. Estimation of voice source waveforms and extraction of MFCCs

This section describes first the two signal processing methods, the quasi-closed phase (QCP) glottal inverse filtering method [36] and the zero frequency filtering (ZFF) method [37]), that are used in the current study for the estimation of the voice source waveforms. After this, the extraction of MFCCs from the two voice source waveforms is described. It is to be noted that source-filter modeling is assumed in QCP but not in ZFF.

2.1. QCP

QCP [36] is a recently proposed glottal inverse filtering method to estimate the glottal waveform from speech. The method is based on the principles of closed phase (CP) [38] analysis which estimates the vocal tract response from a few speech samples located in the closed phase of the glottal cycle using linear prediction analysis. In contrast to the CP method, QCP takes advantage of all the speech samples of the analysis frame in computing the vocal tract model. This is conducted by using weighted linear prediction (WLP) analysis with a specific weighting function called the Attenuated Main Excitation (AME) function [39]. The AME function is designed using glottal closure instants (GCIs) and fundamental period. The AME function attenuates the contribution of the open phase samples in the computation of the speech signal's covariance (or autocorrelation) function. This results in good estimates of the vocal tract transfer function. Finally, the estimate for the glottal flow is obtained by inverse filtering the input speech signal with the vocal tract model. In [36], the accuracy of QCP was shown to be better than that of four existing inverse filtering methods. In addition, the estimation of the glottal flow was shown to be robust with respect to the phonation type. Both of these reasons justify using QCP as a glottal inverse filtering method in the current study. A schematic block diagram describing the steps involved in QCP is shown in Fig. 1.

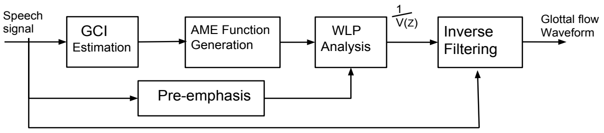

Figure 1: *Block diagram of the QCP method [40].*

2.2. ZFF

ZFF was proposed in [37] based on the idea that the effect of an impulse-like excitation (occuring at glottal closure instant) is present throughout the spectrum including at 0 Hz, while the vocal tract characteristics are mostly reflected at formant peaks at much higher frequencies. In this method, the pre-emphasized speech signal is first passed through a cascade of two zero frequency resonators (pair of poles on the unit circle along the positive real axis in the z-plane). The resulting signal is equivalent to integration (or cumulative sum in the discrete time domain) of the signal four times, hence it grows or decays as a polynomial function of time. The trend is removed by subtracting the local mean computed over the average pitch period (estimated using autocorrelation) at each sample. The resulting output signal is referred as the zero frequency filtered (ZFF) signal. The ZFF signal has an interesting property that its negative-to-positive zero-crossings (NPZCs) correspond to the locations of the impulse-like excitations generated at the instants of glottal closures (GCIs) by considering the positive polarity of the signal [37, 41]. The slope of the ZFF signal at NPZCs provides an estimate of the strength of impulse-like excitation (SoE), which is proportional to the amplitudes of the differentiated electroglottography (EGG) signals at the instants of glottal closure. In this study, we consider the ZFF signal as an approximate voice source waveform as it contains significant information about the voice source. A schematic block diagram describing the steps involved in ZFF is shown in Fig. 2.

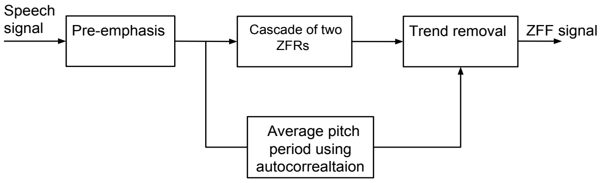

Figure 2: *Block diagram of the ZFF method.*

For an illustration, Fig. 3 shows spectrograms of glottal flow waveforms in three different phonation types estimated using the QCP method. It can be clearly seen that there are large variations in the glottal flow spectra between the three phonation types. In order to capture these variations and to represent them in a compact form, we propose using MFCCs for the source waveforms.

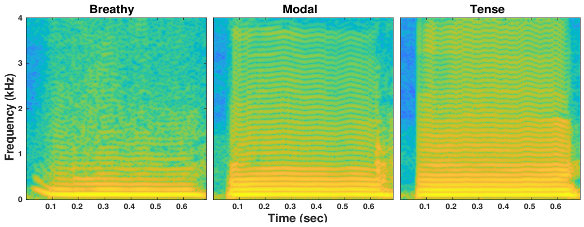

Figure 3: *An illustration of spectrograms of glottal source waveforms estimated using the QCP method for breathy, modal and tense voice.*

2.3. Extraction of MFCCs from voice source waveforms

This section describes the steps involved in deriving MFCCs from the two voice source waveforms obtained using QCP and ZFF. It is to be noted that the proposed feature extraction procedures are similar to the computation of conventional MFCC features from speech, except that the proposed approaches operate on glottal flow waveforms instead of speech signals.

2.3.1. Extraction of MFCCs from the glottal flow estimated using QCP

The schematic block diagram of the extraction of MFCCs from glottal flow waveforms given by QCP is shown in Fig. 4. The procedure consists of mel-band-based analysis of the spectrum of the glottal flow waveform, followed by logarithm and discrete cosine transform (DCT). The spectrum is estimated using a 1024-point FFT with Hamming windowing in 25-ms frames with a 5-ms shift. From the entire cepstrum computed, the first 13 coefficients (including the 0^{th} coefficient) are considered for each frame. The resulting cepstral coefficients (referred as MFCC-QCP) represent the excitation information in a compact form. Also, delta and double-delta coefficients are computed from the static cepstral coefficients.

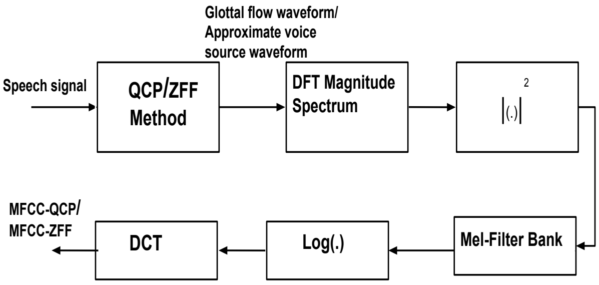

Figure 4: *Extraction of MFCCs from the glottal flow waveform/approximate source waveform estimated using the QCP/ZFF method.*

2.3.2. Extraction of MFCCs from the approximate source waveform computed by ZFF

The schematic block diagram of the MFCC extraction from the approximate voice source waveform estimated by ZFF is shown in Fig. 4. Similarly to the above (sec. 2.3.1), the procedure consists of mel-band-based analysis of the spectrum of the approximate voice source, now estimated with ZFF, followed by logarithm and DCT. The resulting cepstral coefficients are referred as MFCC-ZFF. Here also, static, delta and double-delta coefficients are computed for each frame.

3. Experimental protocol

This section describes the speech database used in the experiments, the reference features that were selected for comparison and the classifier.

3.1. Speech database

In this study, we used a Finnish speech database which consists of 8 different vowels uttered in three phonation types (breathy, modal and tense). The database was collected from 6 female and 5 male speakers whose age ranged between 18–48 years. Each vowel was uttered three times by all the speakers using the three phonation types, making the total number of vowels equal to 792 (3*3*8*11). Speech was recorded in an anechoic chamber at a sampling frequency of 44.1 kHz, but the data was later downsampled to 16 kHz. More details about the database can be found in [21].

3.2. Reference features

Three types of reference features were chosen for comparison: voice quality (VQ) features, conventional MFCCs extracted from speech, and zero-time windowing cepstral coefficients (ZTWCCs). These features were selected based on the findings in [21, 23, 29, 42] for discrimination of speech signals whose voice quality varied from breathy to tense. A brief description of the selected reference features is given below.

3.2.1. Voice quality (VQ) features

The VQ feature set consists of the Normalized Amplitude Quotient (NAQ) [26], Quasi-open quotient (QOQ) [21, 25], H1-H2 [27], Parabolic spectral parameter (PSP) [28] and Maximum dispersion quotient (MDQ) [29]. From these features, NAQ, QOQ, H1-H2 and PSP are extracted from glottal flow signals estimated using inverse filtering and MDQ is derived from wavelet transform on linear prediction residual signal.

3.2.2. Conventional MFCCs

Conventional MFCC features were computed using 25-ms Hamming windowed frames with a 5-ms shift. The first 13 cepstral coefficients (including the 0^{th} coefficient) and their delta and double-delta coefficients were computed yielding a feature vector of 39 elements.

3.2.3. Zero-time windowing cepstral coefficients (ZTWCCs)

Cepstral features derived from high-resolution spectrum obtained by the zero-time windowing (ZTW) method are referred as ZTWCCs. ZTWCCs features were recently proposed for discrimination of phonation types in speech [35]. In this study, the first 13 cepstral coefficients (including the 0^{th} coefficient) derived at the glottal closure instant locations are considered. From static coefficients, delta and double-delta coefficients are computed, which yields a feature vector of 39 elements.

3.3. Classifier

Support vector machines (SVMs) utilizing a radial basis function (RBF) kernel is used for classification [43]. Experiments were conducted with 10-fold cross-validation, where the data was partitioned randomly into 10 equal portions similar [29,35]. One fold was held out to be used for testing with the remaining nine folds for training. Classification accuracies were saved in each fold and this process was repeated for each of the 10-folds. Finally, the mean and standard deviation of the accuracies were considered for evaluation.

4. Results and discussion

Experiments were carried out with the individual feature sets described in Section 2.3 (MFCC-QCP, MFCC-ZFF) and Section 3.2 (VQ, MFCC, ZTWCC) as well as with combinations of these feature sets to analyze the complimentary information between the features. In the combination of features, the complimentary nature of the proposed features with the existing features is focused on.

The combinations considered are: MFCC-QCP+MFCC-ZFF, VQ+MFCC+ZTWCC, VQ+MFCC+ZTWCC+MFCC-QCP, VQ+MFCC+ZTWCC+MFCC-ZFF and combining all.

The results of the experiments for individual features and combination of features are given Table 1 in terms of mean and standard deviation of the classification accuracies. From the table, it can be seen that for the individual feature sets, the proposed MFCC-ZFF features show the highest accuracy (70.54%) compared to the reference features (VQ, MFCCs and ZTWCCs) and to MFCC-QCP. Experiments also revealed that there exists some complimentary information when the MFCC-QCP features are combined with the MFCC-ZFF features. Further, experimentation with the combination of the reference features with the proposed features revealed the existence of complimentary information clearly. Combination of the existing features with the proposed MFCC-ZFF features gave a higher classification accuracy (75.19%) compared to combination of the existing features with the proposed MFCC-QCP (73.54%). The largest improvement in the classification accuracy (76.58%) was achieved when all the features were combined.

Table 1: *Mean and standard deviation of the classification accuracy (in %) after 10-fold cross validation with individual features and combination of features.*

Features	**Mean accuracy**[%]	**Std deviation**[%]
VQ	64.21	4.97
MFCCs	68.52	5.14
ZTWCCs	69.38	4.53
MFCC-QCP	65.23	4.82
MFCC-ZFF	**70.54**	4.15
MFCC-QCP+MFCC-ZFF	71.28	5.13
VQ+MFCCs+ZTWCCs	72.57	3.98
VQ+MFCCs+ZTWCCs+MFCC-QCP	**73.54**	4.67
VQ+MFCCs+ZTWCCs+MFCC-ZFF	**75.19**	4.43
Combining all features	**76.58**	3.96

Table 2 shows the confusion matrix for the combination of all the existing features (i.e., VQ+MFCCs+ZTWCCs). It can be observed that modal phonation is confused with breathy and tense voice. This happens also for the combination of the proposed MFCC-QCP (Table 3) and MFCC-ZFF (Table 4) features with existing features, even though there exists an improvement in overall accuracy and lesser improvement in the accuracy for modal phonation. Major improvement in accuracy with MFCC-ZFF (Table 4) features comes mainly with improvement in detection of breathy voice. From the combination of all features (Table 5), it can be clearly seen that there exists an improvement in the classification accuracy for all the classes and especially for modal phonation. Even though there is an improvement in accuracy, still there exists a confusion between modal and tense phonation. This trend in the accuracies is in line with the previous studies reported in [23,29]. Further, features that capture the production variations of tense phonation are required to reduce the confusion between modal and tense voices, and to improve the overall accuracy.

Table 2: *Confusion matrix (in %) with 10-fold cross validation after combining all the existing features (i.e., VQ+MFCCs+ZTWCCs).*

	Breathy [%]	Modal [%]	Tense [%]
Breathy	76.84	21.15	2.01
Modal	19.68	61.53	18.79
Tense	2.21	18.45	79.34

Table 3: *Confusion matrix (in %) with 10-fold cross validation after combining MFCC-QCP and all the existing features (i.e., VQ+MFCCs+ZTWCCs+MFCC-QCP).*

	Breathy [%]	Modal [%]	Tense [%]
Breathy	79.56	18.01	2.43
Modal	15.28	64.34	20.38
Tense	2.22	21.11	76.67

Table 4: *Confusion matrix (in %) with 10-fold cross validation after combining MFCC-ZFF and all the existing features (i.e., VQ+MFCCs+ZTWCCs+MFCC-ZFF).*

	Breathy [%]	Modal [%]	Tense [%]
Breathy	83.17	13.74	3.09
Modal	13.98	64.30	21.72
Tense	3.79	17.80	78.41

Table 5: *Confusion matrix (in %) with 10-fold cross validation after combining all the features (i.e., VQ+MFCCs+ZTWCCs+MFCC-QCP+MFCC-ZFF).*

	Breathy [%]	Modal [%]	Tense [%]
Breathy	80.89	17.08	2.03
Modal	14.83	68.62	16.55
Tense	1.58	16.93	81.49

5. Summary and conclusions

In this paper, two new feature sets (MFCC-QCP and MFCC-ZFF) were derived from voice source waveforms to classify phonation types in speech. The MFCC-QCP features are derived from glottal flow waveforms estimated with the QCP glottal inverse filtering method and the MFCC-ZFF features are derived from approximate voice source waveforms obtained with the ZFF method. Experiments showed that on its own the MFCC-ZFF features can be used to achieve the highest accuracy in classification. In addition, the results revealed that the proposed MFCC features derived from voice source waveforms provide complimentary information that is present in the existing voice quality parameters, MFCCs and ZTWCCs.

6. Acknowledgements

This study was partly funded by the Academy of Finland (project no. 312490).

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] 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.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

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

[19] E. Shriberg, M. Graciarena, M. Bratt, A. Kathol, S. Kajarekar, H. Jameel, C. Richey, and F. Goodman, "Effects of vocal effort and speaking style on text-independent speaker verification," in *INTERSPEECH*, 2008, pp. 609–612.

[20] 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.

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

[22] 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.

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

[24] 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.

[25] 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.

[26] 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.

[27] I. R. Titze and J. Sundberg, "Vocal intensity in speakers and singers," *the Journal of the Acoustical Society of America*, vol. 91, no. 5, pp. 2936–2946, 1992.

[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] 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.

[30] 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.

[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] G. d. Krom, "A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals," *Journal of Speech, Language, and Hearing Research*, vol. 36, no. 2, pp. 254–266, 1993.

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

[34] M. Borsky, D. D. Mehta, J. H. Van Stan, and J. Gudnason, "Modal and nonmodal voice quality classification using acoustic and electroglottographic features," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 12, pp. 2281–2291, 2017.

[35] S. R. Kadiri and B. Yegnanarayana, "Breathy to tense voice discrimination using zero-time windowing cepstral coefficients (ztwccs)," in *INTERSPEECH*, 2018, pp. 232–236.

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

[37] 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.

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

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

[40] M. Airaksinen, L. Juvela, B. Bollepalli, J. Yamagishi, and P. Alku, "A comparison between straight, glottal, and sinusoidal vocoding in statistical parametric speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1658–1670, 2018.

[41] S. R. Kadiri and B. Yegnanarayana, "Speech polarity detection using strength of impulse-like excitation extracted from speech epochs," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5610–5614.

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

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