Analysis and Detection of Phonation Modes in Singing Voice using Excitation Source Features and Single Frequency Filtering Cepstral Coefficients (SFFCC)

by

sudarsanareddy.kadiri@research.iiit.ac.in kadiri, B Yegnanarayana

in

Annual Conference of the International Speech Communication Association 2018
(*INTERSPEECH-2018*)

Hyderabad, India

Report No: IIIT/TR/2018/-1

Centre for Language Technologies Research Centre
International Institute of Information Technology
Hyderabad - 500 032, INDIA
September 2018

Interspeech 2018
2-6 September 2018, Hyderabad

Analysis and Detection of Phonation Modes in Singing Voice using Excitation Source Features and Single Frequency Filtering Cepstral Coefficients (SFFCC)

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 study, classification of the phonation modes in singing voice is carried out. Phonation modes in singing voice can be described using four categories: breathy, neutral, flow and pressed phonations. Previous studies on the classification of phonation modes use voice quality features derived from inverse filtering which lack in accuracy. This is due to difficulty in deriving the excitation source features using inverse filtering from singing voice. We propose to use the excitation source features that are derived directly from the signal. It is known that, the characteristics of the excitation source vary in different phonation types due to the vibration of the vocal folds together with the respiratory effort (lungs effort). In the present study, we are exploring excitation source features derived from the modified zero frequency filtering (ZFF) method. Apart from excitation source features, we also explore cepstral coefficients derived from single frequency filtering (SFF) method for the analysis and classification of phonation types in singing voice.

Index Terms: Singing voice, Excitation source, Phonation type.

1. Introduction

Humans have the capability of producing a wide variety of variations in voice by manipulating the interaction between vocal folds and vocal tract. This result in the characteristics of voice such as emotions, phonations etc. It is the singing voice that gives an identity to music by providing the meaning that no other instrument can give. One of the most salient features of singing voice is the voice quality/phonation mode. The voice quality is roughly considered as timbre or coloring to the voice. Singer identity and the feelings are expressed through modulations of the voice quality. In this study, our focus is on analysis and classification of phonation modes in singing voice.

According to studies in [1–3], there exists four phonation modes in singing. They are: breathy, modal, flow (or resonant) and pressed phonations. For example, breathy voice may be used to express sweetness, pressed voice may be used for expressing stronger expressions and flow phonation may be encountered in very active singing. The source for phonation modes primarily arises due to the adjustments made at the larynx. From the studies on speech and singing [4, 5], phonation modes have three dimensions: 1) pitch, 2) loudness and 3) laryngeal adjustments. The main production characteristics of four phonation modes are given briefly below.

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 pressed voice contains richer harmonic content [6]. In breathy phonation, there is a reduced vocal fold adduction and minimal vocal fold contact area. This results laxed voice with high level of turbulent noise. Harmonic to noise ratio is generally higher than other phonations [7]. Strong perceptual indicator of breathiness is the sensation of excessive laryngeal airflow [8]. Flow voice is typically produced by a lowered larynx and it is defined more as a vocal technique as it is used exclusively in singing unlike the other modes which require vocal training [9]. The loudness is the key thing in this type where the aim is to achieve higher levels of loudness with lesser effort. The characteristics in this phonation are formant tuning, ample harmonic content and narrowing of the laryngeal vestibule. In modal voice (normal phonation), we can find a full vibration of the vocal folds, along their entire length.

Automatic detection of the phonation mode could help to diagnose vocal disorders such as the hypo-function and hyperfunction of the glottis [10]. Also, many singing students exhibit varying degrees of these malfunctions throughout the course of their studies, teachers could be assisted to correct this during lessons. Apart from singing, phonation modes also play an important role in speech such as for emotion recognition [11, 12].

Several studies have investigated phonation modes from singing with voice quality features derived from inverse filtering [1, 6, 10]. The voice quality feature set consists of normalized amplitude quotient (NAQ) [13], quasi-open quotient (QOQ) [14,15], H1-H2, parabolic spectral parameter (PSP) [16] and maximum dispersion quotient (MDQ) [17]. Subglottal pressure was also found to correlate with the amount of pressedness [18]. The NAQ describes the glottal closing phase and was shown to be robust than the closing quotient when separating breathy, neutral and pressed spoken vowels [15, 17, 19]. This capability transfers to the singing voice. Other features have been proposed for discriminating breathy from tense voices, such as the peak slope [20] and the maxima dispersion quotient (MDQ) [17]. The cepstral peak prominence (CPP) feature was shown to correlate strongly with ratings of perceived breathiness [21]. It was observed that, the voice quality features alone are not sufficient for classification. This is mainly due to inverse filtering problems from singing voice, as singing voice has significant source-filter coupling. In [10], authors used large number of spectral statistics such as spectral centroid, spectral flux, spectral energies in different bands along with various voice quality features and MFCCs. Recently in [6], authors studied the features such as harmonic amplitudes, formant frequencies, bandwidths and amplitudes, harmonic-to-noise ratio along with voice quality features. It was observed that the confusions are between breathy and modal, and flow and pressed phonations.

In this study, we propose to use excitation source features derived directly from the speech signal without using inverse filtering of speech. For deriving these features, we use modified zero frequency filtering (ZFF) method. We also propose to use cepstral coefficients derived from recently proposed single frequency filtering (SFF) method, which provides higher spectro-

10.21437/Interspeech.2018-2502

temporal resolution.

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

2. Signal Processing Methods for Feature Extraction

In this section, we describe features related to excitation source component which are derived from modified ZFF method for singing voice [22]. Also, we describe the features that reflect the effect of excitation in the spectral characteristics, which are derived from the SFF method [23, 24]. It is to be noted that, these two signal processing methods do not assume source-filter model of speech production mechanism.

2.1. Modified zero frequency filtering (ZFF) method

The zero frequency filtering (ZFF) [25] method gives the robust estimates of glottal closure instants (GCIs). The modified ZFF method can handle rapid and wider variations in pitch like in singing voice [22]. In the modified ZFF method, the differenced speech signal is passed through a resonator (given in Eqn. (1)), and the trend in the resonator output ($y_0[n]$) is removed by using a moving average filter (given in Eqn. (2)) [22].

$$y_0[n] = -\sum_{k=1}^{2} a_k y_0[n-k] + x[n], \tag{1}$$

where $a_1 = -2$ and $a_2 = 1$.

$$y[n] = y_0[n] - \frac{1}{2N+1} \sum_{m=-N}^{N} y_0[n+m], \tag{2}$$

where $2N+1$ corresponds to the number of samples used for computing the trend. This operation is repeated twice i.e., passing the signal ($y[n]$) through a resonator and removing the trend. This repetition operation is different from passing the signal through three resonators and removing the trend. The resulting signal oscillates according to variation of local pitch period [22], and is referred to as modified ZFF signal. The instants of negative-to-positive zero crossings (NPZCs) correspond to the glottal closure instants (GCIs).

Features of excitation source: From the modified ZFF method, we derive the excitation features such as strength of excitation (SoE), energy of excitation (EoE), loudness measure and ZFF signal energy. Let the epochs be denoted by $\mathcal{E} = \{e_1, e_2, ..., e_M\}$, where M is the number of epochs. The time duration between two successive epochs gives the instantaneous fundamental period (or the pitch period T_0), and its reciprocal gives the instantaneous fundamental frequency (F_0).

The slope of the ZFF signal around each NPZCs corresponds to the SoE, which is proportional to the rate of closure of the vocal folds [26]. A measure of SoE around the GCI is given by

$$SoE = |y[e_k + 1] - y[e_k - 1]|, \qquad k = 1, 2, ..., M. \tag{3}$$

The energy of excitation (EoE) feature is computed from the samples of the Hilbert envelope of the LP residual over 2 ms region around each GCI. This measure gives the vocal effort [26]. A 10^{th} order LP analysis is used for each frame of 16 ms and a frame shift of 2 ms.

$$EoE = \frac{1}{2K+1} \sum_{i=-K}^{K} h_e^2[i], \tag{4}$$

where 2K+1 corresponds to the number of samples in the 1 ms window. Loudness (perceived loudness) measure captures the abruptness of glottal closure [27] and it is 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 and is computed as

$$v_{zff}[n] = \frac{1}{L} \sum_{i=-L/2}^{L/2} y^2[n+i], \tag{5}$$

where $y[n]$ is the ZFF signal, and L corresponds to the window length (10 ms) over which the energy is computed. The energy of the ZFF signal at GCI is considered in this study. These features are shown to be useful for the analysis and discrimination of phonations and emotions in speech [28, 29].

2.2. SFF method and Extraction of SFFCC

The objective of SFF is to derive the amplitude envelope of the signal as a function of time. The spectro-temporal resolution can be adjusted by varying the r parameter, which represents the pole location in the z-plane. The steps involved in SFF method are as follows [23, 24].

1. The input speech signal $s[n]$ is pre-emphasized to remove any low frequency components.
$$x[n] = s[n] - s[n-1]. \tag{6}$$
2. The signal ($x[n]$) is multiplied with a complex exponential $e^{j\bar{w}_k n}$, where $\bar{w}_k = \pi - w_k = \pi - \frac{2\pi f_k}{f_s}$. The resulting frequency shifted signal is represented by
$$x[n,k] = x[n]e^{j\bar{w}_k n}, \tag{7}$$
where k ranges from 0 to K, (K=$f_s/2$).
3. The frequency shifted signal $x[n,k]$ is passed through a single-pole filter $H(z)$, where
$$H(z) = \frac{1}{1 + rz^{-1}}. \tag{8}$$
Here, the r value is chosen as 0.995.
4. The output of the filter ($y[n,k]$) is given by
$$y[n,k] = -ry[n-1,k] + x[n,k]. \tag{9}$$
The amplitude envelope of the signal $y[n,k]$ is given by
$$v[n,k] = \sqrt{(y_r[n,k])^2 + (y_i[n,k])^2}, \tag{10}$$
where y_r, y_i represents the real and imaginary parts respectively. The term $v[n,k]$ corresponds to the SFF envelope of the signal at frequency f_k. The magnitude spectrum can be obtained for each instant of n.

Figure 1 gives an illustration of SFF spectrograms for breathy, modal, flow and tense/pressed phonations for vowel A in Soprano singing category. It can be clearly seen that there exists a significant variations in the spectrum. In order to capture these variations, we propose to derive the single frequency filtering cepstral coefficients (SFFCCs).

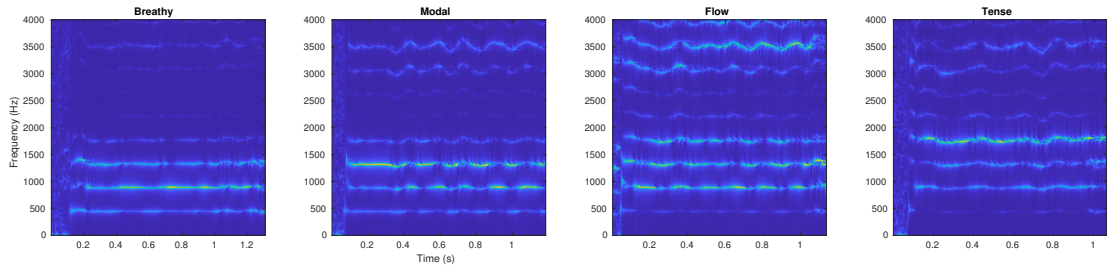

Figure 1: *An illustration of SFF spectrograms for breathy, modal, flow and tense phonations for vowel A in Soprano singing category.*

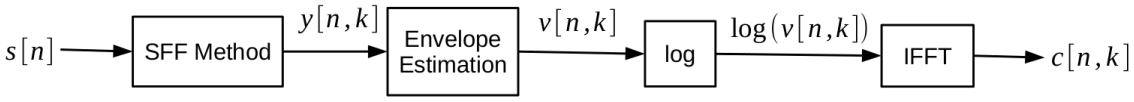

Figure 2: *Block diagram of single frequency filter cepstral coefficients (SFFCCs) extraction [30].*

2.2.1. SFFCC Extraction

Cepstrum $c[n,k]$ is computed from SFF spectrum $v[n,k]$, and is given by

$$c[n,k] = IFFT(log(v[n,k])). \tag{11}$$

From $c[n,k]$, first 13 cepstral coefficients are considered and they are named as single frequency filtering cepstral coefficients (SFFCCs). The SFFCCs can be obtained at each sampling instant. In this study, instead of computing at each instant, we computed the SFFCCs at GCI locations. The schematic block diagram of SFFCCs extraction is shown in Fig. 2.

2.3. Feature analysis

The distributions of the proposed excitation source features (SoE, EoE, Loudness and ZFF energy) are given in Fig. 2. It can be seen that SoE values are high for breathy, and low for pressed and flow phonations. EoE values are high for pressed voice, followed by flow, modal and breathy voice (low EoE values). This parameter indicates the vocal effort required for producing the phonation type. The perceived loudness values are low for breathy voice than for modal, and pressed and flow phonations have relatively higher values. The ZFF signal energy comes out to be lower for pressed and flow followed by modal and breathy. From the box plots, it can also be observed that there exists a significant overlap of the feature values between modal and breathy, and flow and tense voices. This is also confirms the studies reported in [1, 6, 10]. In all, there exists a good separation among the feature values for the discrimination of phonation types.

3. Experimental protocol

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

3.1. Database used

We use the phonation dataset described in [1], which contains sustained vowels sung by a soprano female professional (in singer's native language, Russian) recorded at a sampling frequency of 44.1 KHz. The phonation modes correspond to Sundberg's definitions of breathy, neutral, flow and pressed voice [3].

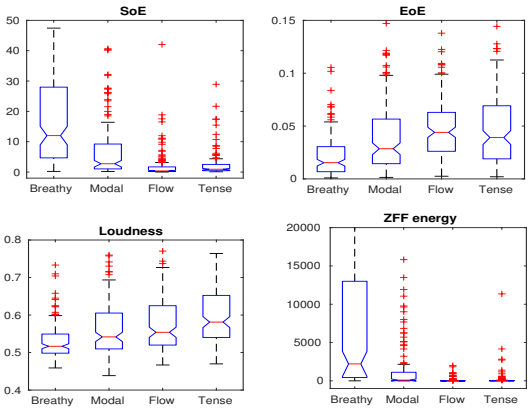

Figure 3: *Distribution of features for breathy, modal, flow and tense phonation types using box plots. The central mark indicates the median, and the bottom and top edges of the box indicate the 25^{th} and 75^{th} percentiles, respectively. The whiskers extend to the most extreme data points not considered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

This database consists of 763 recordings of nine different vowels: A, AE, I, O, U, UE, Y, OE and E and with pitches ranging from A3 to G5 [1]. The mean duration of the recorded samples is 1.3 seconds with variation from 0.9 to 1.6 seconds. The full dataset is used to enable a comparison with classification results from previous studies.

3.2. Features for comparison

MFCCs and voice quality features are considered for comparison. These features are selected based on the findings in [15, 17, 19], which showed the most suitable features for discrimination of phonation modes. The voice quality feature set consists of NAQ, QOQ, H1-H2, PSP, and MDQ. Out of these, first four features are derived from Inverse Filtering method [31]. A brief description of the features is given below.

3.2.1. Normalized Amplitude Quotient (NAQ)

NAQ [13] is computed from two amplitude values, and it is given by

$$NAQ = \frac{f_{AC}}{d_{min} \cdot T}, \tag{12}$$

where, f_{AC} is the AC-amplitude of the flow, d_{min} is the negative peak amplitude of the glottal flow derivative and T is the pitch period.

3.2.2. H1-H2

H1-H2 [15], is the difference between the amplitudes of the first two harmonics in the voice source spectrum.

3.2.3. Quasi-open quotient (QOQ)

The QOQ [14, 15] is related to the open quotient. Quasi-open phase divided by the local pitch period gives QOQ.

3.2.4. Maximum dispersion quotient (MDQ)

The MDQ [17] parameter measures the dispersion in the LP residual around the GCI and it captures abruptness of glottal closure.

3.2.5. Parabolic Spectral Parameter (PSP)

PSP [16] is derived by fitting a parabola to the low frequency of glottal flow spectrum.

3.2.6. MFCCs

In this study, we derived 13 mel-frequency cepstral coefficients using 25 ms Hamming windowed frames, with 5 ms shift.

4. Classification Experiments and Discussion on Results

The experiments are carried out using support vector machines (SVMs) with a radial basis function (RBF) kernel [32]. Classification experiments are conducted using 10-fold cross-validation. The dataset is partitioned (randomly) into 10 equal sets and one fold is held out to be used for testing with the remaining set for training. In each fold classification accuracies are saved and the process is repeated for 10-folds. The experiments are carried out for 7 different feature vectors:

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

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

Features	**Mean accuracy**[%]	**Standard deviation**[%]
VQ	29.18	10.17
MFCC	61.05	06.33
VQ+MFCC	34.47	15.32
Excitation	52.11	05.92
SFFCC	65.24	04.05
Excitation+SFFCC	67.12	06.12
MFCC+Excitation+SFFCC	**70.92**	06.24

The results of the 10-fold cross validation experiment are shown in terms of mean and standard deviation of the classification accuracies in Table 1. From the table, it can be seen that including parameters such as MFCCs, excitation features, and SFFCCs (i.e., MFCC+Excitation+SFFCCs) gives the highest average classification accuracy (70.92%). It can also be observed that classification accuracy with SFFCCs gives the highest compared to VQ features and MFCCs. It is to be noted that, voice quality features are not able to to discriminate phonation modes in singing unlike in speech [17, 19]. This is mainly because of the inverse filtering issues, as most of the features (except MDQ) are derived from the glottal flow waveform. It is known that inverse filtering methods fail for high pitched voices [14, 33]. The four proposed excitation source features (directly estimated from speech signal) are giving 52.11%. Combining the excitation features with SFFCCs increases the accuracy (67.12%). This indicates that the excitation features and SFFCCs are providing complimentary information. The best classification accuracy of 70.92 % is achieved when the proposed features are combined with the MFCCs.

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

	Breathy [%]	Modal [%]	Flow [%]	Tense [%]
Breathy	73.54	25.40	0	1.06
Modal	16.91	65.44	4.41	13.24
Flow	0	3.96	58.42	37.62
Tense	2.10	10.81	20.72	66.37

Table 3: *Confusion matrix (in %) with 10-fold cross validation after combining MFCC, Excitation and SFFCC features (i.e., MFCC+Excitation+SFFCC).*

	Breathy [%]	Modal [%]	Flow [%]	Tense [%]
Breathy	83.77	15.71	0	0.52
Modal	13.07	72.55	3.27	11.11
Flow	0	1.79	57.14	41.07
Tense	1.37	5.92	28.62	64.19

Table 2 shows the confusion matrix for the combination of excitation source features and SFFCCs (i.e., Excitation+SFFCCs). From the table, it can be observed that there is a significant confusion between breathy and modal voice, and flow and tense voice. This observation also complies with the results reported in [1, 6, 10]. Table 3 shows the confusion matrix for the combination of MFCCs, excitation source features and SFFCCs (i.e., MFCC+Excitation+SFFCCs). It can be observed that there is a significant reduction in confusion between breathy and modal voice. But there is not much reduction in discrimination between flow and tense voice. The discrimination can be further improved by including other voice quality parameters. Also, there is a need for exploring features that can capture the effect of excitation on the vocal tract system [3, 5], especially for the discrimination of breathy and modal, and flow and tense voices.

5. Summary and conclusion

In this paper, we investigated the discriminative and explanatory power of excitation source features derived from modified ZFF method and cepstral coefficients derived from SFF method for phonation mode classification. From the experimental results, it was shown that proposed excitation features and SFFCCs provides better discrimination of phonation modes. The existing voice quality features lack in accuracy, mainly because of the inverse filtering issues especially for high pitched voices like singing. On the other hand, proposed features do not use source-filter model of speech production for deriving features. This suggests that the proposed features can be useful for analyzing continuous speech/singing.

6. References

[1] P. Proutskova, C. Rhodes, T. Crawford, and G. Wiggins, "Breathy, resonant, pressed–automatic detection of phonation mode from audio recordings of singing," *Journal of New Music Research*, vol. 42, no. 2, pp. 171–186, 2013.

[2] J. Sundberg, "The perception of singing," in *The Psychology of Music (Second Edition)*. Elsevier, 1999, pp. 171–214.

[3] J. Sundberg", *The science of the singing voice*. Illinois University Press, 1987.

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

[5] J. Sundberg, "The acoustics of the singing voice," *Scientific American*, vol. 236, pp. 82–91, 1977.

[6] J.-L. Rouas and L. Ioannidis, "Automatic Classification of Phonation Modes in Singing Voice: Towards Singing Style Characterisation and Application to Ethnomusicological Recordings," in *interspeech*, vol. 2016, 2016, pp. 150 – 154.

[7] 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.

[8] E. U. Grillo and K. Verdolini, "Evidence for distinguishing pressed, normal, resonant, and breathy voice qualities by laryngeal resistance and vocal efficiency in vocally trained subjects," *Journal of Voice*, vol. 22, no. 5, pp. 546–552, 2008.

[9] J. Sundberg, "Vocal fold vibration patterns and modes of phonation," *Folia phoniatrica et logopaedica*, vol. 47, no. 4, pp. 218–228, 1995.

[10] D. Stoller, S. Dixon *et al.*, "Analysis and classification of phonation modes in singing," 2016.

[11] 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.

[12] S. Patel, K. R. Scherer, E. Bjrkner, and J. Sundberg, "Mapping emotions into acoustic space: The role of voice production," *Biological Psychology*, vol. 87, no. 1, pp. 93 – 98, 2011.

[13] 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.

[14] T. Drugman, P. Alku, A. Alwan, and B. Yegnanarayana, "Glottal source processing: From analysis to applications," *Computer Speech & Language*, 2014.

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

[16] 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.

[17] 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.

[18] M. Millgård, T. Fors, and J. Sundberg, "Flow glottogram characteristics and perceived degree of phonatory pressedness," *Journal of Voice*, vol. 30, no. 3, pp. 287–292, 2016.

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

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

[21] 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.

[22] S. R. Kadiri and B. Yegnanarayana, "Analysis of singing voice for epoch extraction using zero frequency filtering method," in *ICASSP*, Apr. 2015, pp. 4260–4264.

[23] G. Aneeja and B. Yegnanarayana, "Single frequency filtering approach for discriminating speech and nonspeech," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 23, no. 4, pp. 705–717, Apr. 2015.

[24] S. R. Kadiri and B. Yegnanarayana, "Epoch extraction from emotional speech using single frequency filtering approach," *Speech Communication*, vol. 86, pp. 52 – 63, 2017.

[25] 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.

[26] 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.

[27] 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.

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

[29] S. R. Kadiri and B. Yegnanarayana, "Breathy to tense voice discrimination using zero-time windowing cepstral coefficients (ZTWCCs)," in *INTERSPEECH*, Sept. 2018.

[30] K. N. R. K. R. Alluri, S. Achanta, S. R. Kadiri, S. V. Gangashetty, and A. K. Vuppala, "SFF anti-spoofer: IIIT-H submission for automatic speaker verification spoofing and countermeasures challenge 2017," in *Interspeech 2017*, 2017, pp. 107–111.

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

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

[33] 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.