Analysis of Breathy Voice based on Excitation Characteristics of Speech Production

Sathya Adithya Thati, Bajibabu Bollepalli, Peri Bhaskararao and B. Yegnanarayana
Speech and Vision Lab,
International Institute of Information Technology, Hyderabad, India
Email: sathya.adithya@research.iiit.ac.in, bajibabu.b@research.iiit.ac.in, bha.peri@iiit.ac.in and yegna@iiit.ac.in

***Abstract*—The objective of this paper is to find the fundamental difference between breathy and modal voices based on differences in speech production as reflected in the signal. We propose signal processing methods for analyzing the phonation in breathy voice. These methods include technique of zero-frequency filtering, loudness measurement, computation of periodic to aperiodic energy ratio and extraction of formants and their amplitudes using group-delay based technique. Parameters derived using these methods capture the excitation source characteristics which play a prominent role in deciding the voice quality. Classification of vowels into breathy or modal voice is achieved with an accuracy of 93.93% using the loudness measure.**

***Index Terms*—Breathy voice, modal voice, zero-frequency filtering, PAP, excitation characteristics, loudness measure**

I. Introduction

Modal voice, breathy voice and creaky voice are some of the major ‘phonation types’. Phonation refers to the manner in which vocal folds vibrate. Phonation types are also known as ‘states of glottis’. Glottis is the area between the vocal folds. Breathy voice contrasts with modal voice in some languages such as Gujarati, Marathi and Hindi. While modal voice is produced by ‘regular vibrations of the vocal folds’ at any frequency within the speaker’s normal range, in breathy voice vocal folds vibrate ‘without appreciable contact’ with arytenoid cartilages further apart than in modal voice and with ‘higher rate of airflow than in modal voice’ [1].

Gujarati is one of the few languages known for distinguishing breathy and modal phonation in both consonants and vowels, as in the words: bAr meaning ‘twelve’; b:Ar meaning ‘burden’; and bA:r meaning ‘outside’ – where b: is a breathy voiced consonant and A: is a breathy voiced vowel. ‘:’ is used as a notation to represent the breathy variant of a vowel or a consonant in this paper. Apart from its linguistic function as in Gujarati and other languages, breathy voice also characterizes inherent voice quality such as in husky voice and also in pathological speech as in Parkinson’s disease, dysphonia and dysarthria [2], [3] where it is believed to be due to glottal leakage of air.

From an articulatory perspective, breathy voice is a different type of phonation from aspiration. However, breathy-voiced and aspirated stops are acoustically similar in that in both cases there is an audible period of breathiness following the stop.

The difference in the production of a breathy voice and a modal voice lies in the way the vocal tract system is excited. Breathy voice consists of a more open laryngeal configuration compared to modal voice. Thus there is variation in the way the vocal tract system is excited. The excitation source plays a prominent role in production of speech for deciding the voice quality.

In the literature, acoustic measures of breathy voice are not often explicitly described. The analysis of breathy voice was mostly influenced by the usual spectrum analysis and spectrographic methods. A few source features were measured through spectrum analysis. This involved computation of features such as fundamental frequency (F0), formant frequencies, acoustic intensity, periodicity, additive noise and spectral tilt [4]. H1-H2, H1-A1, H1-A2 and H1-A3 [5], [6], [7] and Normalized amplitude quotient (NAQ) of the glottal waveform and its derivative waveform characterize the spectral slope properties of the breathy voice [8]. Here H1 and H2 are the amplitudes of the first and second harmonics, respectively and A1, A2, A3 are the amplitudes of the first, second and third formant frequencies, respectively. Glottal to noise excitation ratio (GNE) [9], [10] and Harmonics to noise ratio (HNR) reflect the presence of aspiration noise components in breathy voice. F1F3syn, which is a synchronization measure between the amplitude envelopes of the first and third formant frequency band signals is reported in [11] and NBP (Normalized breathiness power measure), which is calculated based on F1F3syn is used to characterize the amount of breathiness present in a signal [12]. Jitter, shimmer and higher order statistics (HOS) properties (like skewness and kurtosis of the data samples) were computed in [13]. The higher order statistics could not be attributed to the speech production mechanism involved in the case of breathy voice. A few new measures like harmonic energy of residue harmonic to signal ratio and number of voiced frames were also introduced in [14] for characterizing breathiness.

Analysis based on excitation source attempts to take into account the timing information of the glottal activity. In this paper, we propose features based on source characteristics and robust techniques for analysis of breathy voices.

The paper is organized as follows. In Section 2, the speech production mechanism underlying the breathy phonation is explained. Section 3 explains the data used for the analysis. In

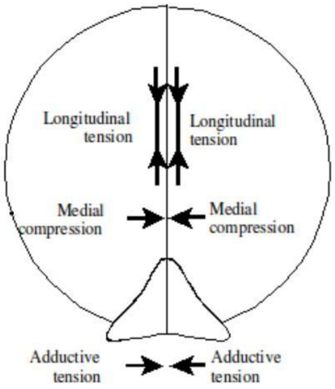

Fig. 1. *Laryngeal parameters in the articulatory description of phonation types [16].*

Section 4, methods used for analysis are explained. In Section 5, results of the analysis are discussed. Finally, summary and conclusions are presented in Section 6.

II. Production mechanism of breathy voices

Human beings can produce speech sounds with not only regular voicing vibrations at a range of different pitch frequencies, but also with a variety of voice source characteristics reflecting different voice qualities.

The voice is controlled by different types of muscular tensions, namely, adductive tension, medial compression and longitudinal tension [15]. Adductive tension is controlled by interarytenoid muscles and draws the arytenoids together. Medial compression is controlled by the lateral cricoarytenoid muscles and keeps the ligamental glottis closed. Similarly, longitudinal tension is mediated primarily by muscles of the vocal folds and the cricothyroid muscles. The contraction of the cricoarytenoid muscles can also increase the longitudinal tension by tilting the arytenoid cartilages backwards. Figure 1 shows the laryngeal parameters in the articulatory description of phonation types [16].

Breathy voice can be produced by maintaining an open glottis for the majority or entirety of the vibration cycle, or it can be caused by vocal folds which close more slowly than for modal phonation. Breathiness is characterized by low adductive tension and moderate to high medial compression. A triangular opening between the arytenoid cartilages, which is a consequence of the low adductive tension combined with the configuration of voicing, produces a breathy voiced phonation. It is characterized by vocal folds having little longitudinal tension. This results in some turbulent air flow through the glottis. Thus we have the auditory impression of "voice mixed in with breath" [5].

Breathy voice and whisper are different. They differ in the manner of production. In breathy voice the vocal muscle tension is low and there is voicing, whereas in whisper the vocal muscle tension is high and there is no voicing involved.

III. Data collection

Initial data for the analysis was collected at 48kHz sampling frequency in a quiet room, and was recorded by an expert phonetician. Data was also recorded from 10 native Gujarati speakers (8 male and 2 female). The participants include those who have stayed most of the time in Gujarat, and those who haven't been there but are natives of that language.

Syllables containing the breathy voice phonation in the vowel part were recorded by an expert phonetician. Some words containing the breathy phonation and their corresponding contrasting modal phonated words were recorded by the native Gujarati speakers. To ensure uniform prosodic effect, and to help the speakers to speak naturally, meaningful declarative carrier words and sentences were used.

IV. Parameters/features for breathy voice

Following are the techniques used to compute the parameters for the analysis of the breathy voiced signals. These techniques are robust because they attempt to capture the acoustic properties of the actual speech production mechanism.

A. Zero-frequency filtering technique

A method is proposed for extraction of the instantaneous F0, epoch extraction and strength of impulse-like excitation at epochs [17], [18]. The method uses the zero-frequency filtered signal derived from speech to obtain the epochs (instants of significant excitation of the vocal tract system) and the strength of impulse at the epochs.

The method involves passing the differenced speech signal through a cascade of two ideal digital resonators, each located at 0 Hz. The trend in the output is removed by subtracting the local mean at each sample, computed over a window length in the range of about 1 to 2 pitch periods. The negative to positive zero crossing instants in the resulting zero frequency filtered (ZFF) output are called epochs. The slopes of the ZFF signal at epochs give the relative strengths of the impulse-like excitation (SoE) around epochs. The reciprocal of the interval between successive epochs gives the instantaneous fundamental frequency (F0).

It is observed that the F0 of a speaker is lowered during breathy phonation when compared to F0 during modal voice as shown in Figures 2 and 3. The decrease in the overall fundamental frequency values can be attributed to the fact that during the breathy voice, there is a gap for air to flow through the vocal folds and this results in the slower vibrations of the vocal folds. Breathy phonation starts with lower F0 and increases steeply in a short duration. The rise is as high as 20% in less than a period of 10 milliseconds.

We also observe from Figure 3 that there is a sudden rise in the SoE from the stop consonant to the vowel in the modal voice signal as anticipated, whereas we observe in Figure 2 that the transition in the SoE for breathy voice is gradual. This is because there is no abruptness in the glottal closure mechanism of a breathy phonation.

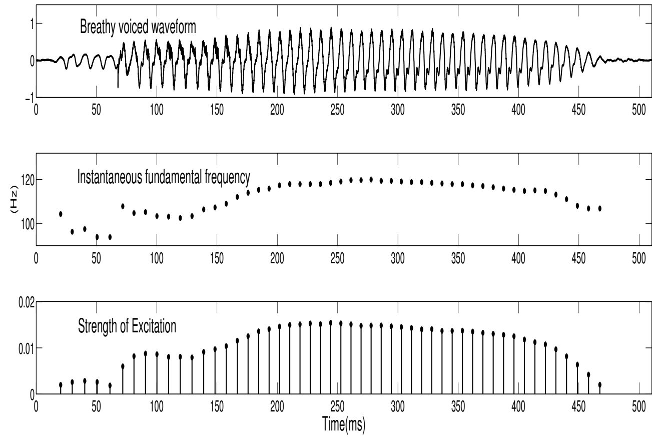

Fig. 2. *Breathy voiced syllable (/bi:/)'s (a) Waveform, (b) Instantaneous fundamental frequency and (c) Strength of Excitation.*

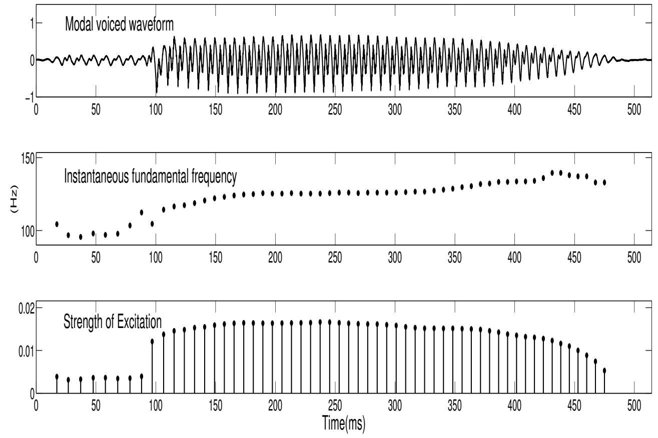

Fig. 3. *Modal voiced syllable (/bi/)'s (a) Waveform, (b) Instantaneou fundamental frequency and (c) Strength of Excitation.*

B. Periodic-aperiodic energy computation

Breathy voice speech consists of increased spectral noise, particularly at higher frequencies. This is due to persistent leakage of air through the glottis during breathy phonation. The ratio of the periodic and aperiodic energies (PAP) can be used as a measure to reflect this property.

This approach to calculate PAP involves iterative decomposition of speech into periodic and aperiodic components as proposed in [19]. The method is summarized in the following steps:

(a) Perform linear prediction (LP) analysis to compute the LP-residual

(b) Divide the LP residual into frames of size 32 ms with a frame shift of 4 ms. Check for voiced and unvoiced frames.

(c) Compute cepstrum using 512 point FFT and Hamming window. Identify the peak in cepstrum relating to harmonics in spectrum by using pitch information obtained by ZFF method (Section 3.1).

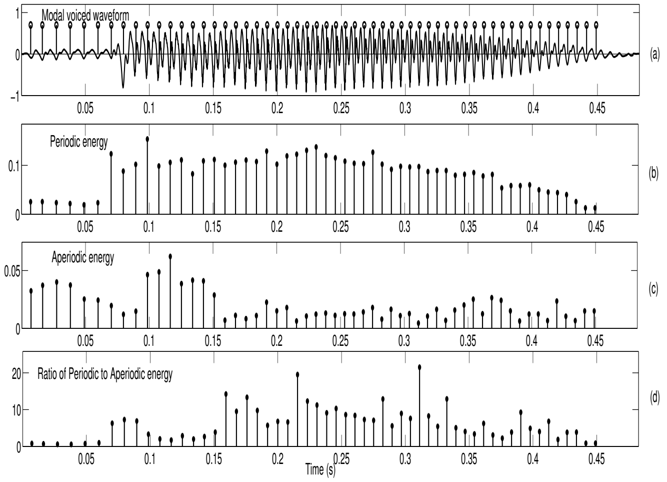

Fig. 4. *PAP for modal voice (/bi/) : (a) Waveform, (b) Periodic energy, (c) Aperiodic energy and (d) PAP ratio.*

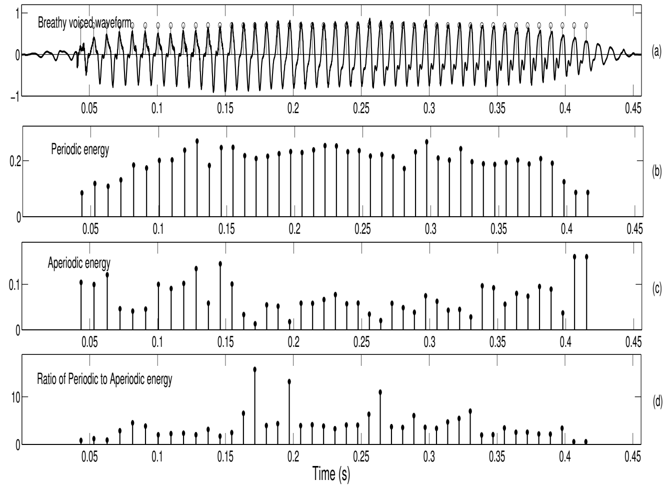

Fig. 5. *PAP for breathy voice (/bi:/) : (a) Waveform, (b) Periodic energy, (c) Aperiodic energy and (d) PAP ratio.*

(d) Compute the harmonic log spectrum by making all the coefficients in cepstrum, except the 9 samples around the peak corresponding to the pitch period to zero, and take IDFT.

(e) Compute the spectrum of the LP residual frame. Samples from the spectrum are now divided into periodic and aperiodic parts.

(f) An iterative algorithm is used to compute the aperiodic component of the residual. Periodic component is obtained by subtracting the aperiodic component from the residual of the speech signal.

(g) Synthesize periodic and aperiodic components of the speech signal by exciting the all pole filter (LP synthesis) with the periodic and aperiodic components of the residual as excitation, respectively.

The ratio of energy of the periodic and aperiodic components ($\frac{E_p}{E_{ap}}$) computed over each of the frames in the voiced

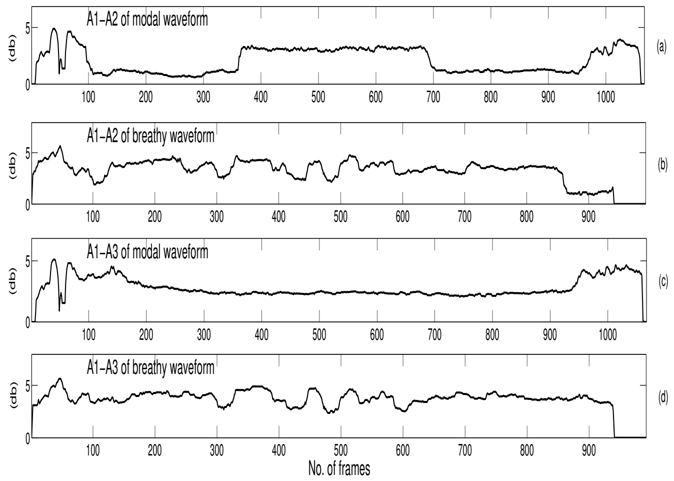

Fig. 6. *Spectral Tilt of modal (/bo/) and breathy (/bo:/) voices : (a) A1-A2 for modal voice, (b) A1-A2 for breathy voice, (c) A1-A3 for modal voice and (d) A1-A3 for breathy voice.*

regions of the utterances is analyzed. Since the intensity of noise is higher at higher frequencies in breathy voice speech signal, we observe that the aperiodic energy is significantly high in breathy signals than that in modal signals, resulting in lesser PAP value for breathy vowels. This is shown in Figures 4 and 5.

C. *Formant extraction method*

Spectral Tilt is a measure of the degree to which intensity drops off as frequency increases. It is one of the acoustic parameters used to differentiate breathy phonation type from other phonation types. It is quantified by comparing the amplitudes of the first harmonic to that of higher frequency harmonics, which could be the second harmonic or the formant frequencies. Spectral tilt is observed to be more for breathy vowels, which means that there is higher fall off in energy at higher frequencies in the signal. The values of the measures used to define the spectral tilt (H1-H2, H1-A1, A1-A2 and A1-A3) are higher for breathy vowels compared to that for its modal counterpart. Locations of formants are computed using the group delay based method given in [20].

The computed H1-H2 values are shown in Table I. A1-A2 and A1-A3 plots of modal and breathy speech are shown in Figure 6.

TABLE I
Table showing the H1, H2 and H1-H2 values for breathy and modal sounds at different instances (in dB).

Breathy voice (/ba:/)			Modal voice (/ba/)		
H1	H2	H1-H2	H1	H2	H1-H2
-17	-30.7	13.7	-20.5	-24.4	3.9
-20.3	-29.3	9	-20.3	-24.4	4.1
-16.6	-28.9	12.3	-22.5	-25	2.5

D. *Measure of loudness*

Perceived loudness of speech is related to the abruptness of the glottal closure. In a breathy voice, the glottal closure is not so abrupt compared to modal voice, and hence the perceived loudness of breathy speech is less compared to the perceived loudness of modal speech. This can be used as a measure to compare different voice qualities.

An objective measure (η) of perceived loudness based on the abruptness of glottal closure derived from the speech signal is discussed in [21]. The abruptness of the glottal closure derived from the EGG signal was shown to be high for loud speech compared to soft and normal speech. When the glottal closure is abrupt, the Hilbert envelope of the LP residual of the speech signal will have sharper peaks at the epochs. The sharpness of the peaks in the Hilbert envelope at the epochs is derived by computing the ratio $\eta = \frac{\sigma}{\mu}$. Here μ denotes the mean, and σ denotes the standard deviation of the samples of the Hilbert envelope of the LP residual in a short interval (2 ms) around the epochs.

Table II shows the means of the loudness values calculated for breathy and modal vowels.

TABLE II
Table showing the mean loudness values for breathy and modal vowels.

Vowel	Breathy	Modal
a	0.51	0.70
e	0.53	0.72
i	0.53	0.69
o	0.53	0.72
u	0.52	0.67

A classification experiment was conducted to classify a speech vowel as either breathy voiced or modal voiced using the derived loudness measure (η). A suitable threshold value was set and all the test samples whose loudness measure was obtained to be less than the threshold were classified as breathy voiced and the remaining as modal voiced. An accuracy of 93.93% is obtained in classifying the test samples. This indicates that the breathy voice quality and the perception of loudness are closely related. The more the breathiness in a speech signal, the less louder we perceive the speech.

V. Results

Duration of the stop consonant preceding the breathy vowel is observed to be lesser than that for the modal voice. This is due to lower vowel onset time for the breathy speech.The reason attributed to this is that the speaker knows before hand that the succeeding phone is breathy, and thus his production mechanism is preset to that of a breathy voice. In this configuration, it is difficult for the speaker to utter the consonant for a longer duration. This initial setting of the production mechanism for the breathy sounds is also one of the reasons behind the gradual transition of SoE in Figure 2.

The mean values of the parameters computed for the breathy vowels and its modal counterpart are given in Table III. In the case of a naturally breathy voice, the contrast between the breathy and modal voice is lesser than the contrast observed for a naturally non-breathy voice.

TABLE III

Table showing the mean values of the parameters for breathy and modal sounds.

Parameter	Breathy	Modal
F0 (Hz)	114	123
SoE	0.0118	0.0101
PAP	9.93	19.00
Loudness	0.52	0.70
A1-A2 (dB)	2.61	2.21
A1-A3 (dB)	3.25	2.84

Both conventional features such as spectral tilt and new features like $PAP, loudness, SoE$ are used to describe the acoustic characteristics of breathy voice quality. These features can be used to spot breathiness in a speech signal. We observe that breathy voice is perceived to be less loud than the modal voice and it is measured by a loudness measure. Due to higher amount of aperiodicity attached with breathy phonation, we see that the PAP ratio is less for breathy voice quality. The average F0 is less and the strength of excitation is more for breathy voice. The spectral tilt is higher for breathy voice as confirmed by the measures of A1-A2 and A1-A3.

VI. Summary and conclusions

In this paper we have emphasized the need to look into features based on excitation characteristics to derive acoustic cues for breathy voices. Features based on instantaneous fundamental frequency contour, the strength of impulse-like excitation at epochs, the ratio of the periodic and aperiodic components of speech and loudness measure derived from the HE of the LP residual were proposed to supplement the spectral features for characterizing breathy voices. The derived loudness measure was used to classify breathy vowels. The studies reported in this paper help to derive features of breathy voice for spotting such segments in continuous speech.

References

[1] Peter Ladefoged and Ian Maddieson, “The Sounds of the World’s Languages”. Oxford: Blackwell. 1996.

[2] J. Hillenbrand and R.A. Houde, “Acoustic Correlates of Breathy Vocal Quality: Dysphonic Voices and Continuous Speech,” J Speech Hear Res, vol. 39, pp. 311-321, Apr. 1996.

[3] Meysam Asgari and Izhak Shafran, “Extracting cues from Speech for predicting severity of Parkinson’s disease”, in International Workshop on Machines Learning for Signal Processing, Kittila, Finland, 2010.

[4] Barbara Blankenship, “The timing of nonmodal phonation in vowels”, Journal of Phonetics, vol. 30, no. 2, pp. 163-191, 2002.

[5] Matthew Gordon and Peter Ladefoged, “Phonation Types: a cross-linguistic overview”, Journal of Phonetics, 29, 383-406, 2001.

[6] Ratree Wayland and Allard Jongman, “Acoustic correlates of breathy and clear vowels: the case of Khmer”, Journal of Phonetics, 31, 181-201, 2003.

[7] Hanson H., “Glottal characteristics of female speakers: Acoustic correlates”, J. Acoustic. Soc. Amer., Vol. 101: 466-481. , 1997

[8] Alku, P., Vilkman, E., “Amplitude domain quotient for characterization of the glottal volume velocity waveform estimated by inverse filtering”, Speech Communication, Vol. 18(2), 131-138, 1996

[9] M. Fröhlich, D. Michaelis, and H. W. Strube, “Acoustic “breathiness measures” in the description of pathological voices,” in Proc. ICASSP, vol. 2, pp. 937-940, Seattle, WA, May 1998.

[10] Michaelis, D., Gramss, T., Strube, H.W., “Glottal-to-noise excitation ratio - a new measure for describing pathological voices”, Acustica, Vol.83, 700-706, 1997

[11] Ishi C.T., Ishiguro H., Hagita N., “Analysis of the roles and the dynamics of breathy and whispery voice qualities in dialogue speech”, EURASIP Journal on Audio, Speech, and Music Processing 2010, ID 528193, 1-12, Jan. 2010.

[12] Ishi C.T., Ishiguro H., Hagita N., “Improved Acoustic Characterization of Breathy and Whispery Voices” In Proc. of Interspeech 2011, pp. 1237-1240.

[13] Ji-Yeoun Lee, Sangbae Jeong, Minsoo Hahn and Hong-Shik Choi, “Automatic voice quality measurement based on efficient combination of multiple features”,in ICBBE, pp. 1272 - 1275 , May 2008.

[14] Eduardo Castillo-Guerra, Adel Ruz, “Automatic modeling of acoustic perception of breathiness in pathological voices”, IEEE Transaction Biomedical Engineering vol.56 (4), pp.932-940, 2009

[15] C. Gobl, “A preliminary study of acoustic voice quality correlates”, STL-QPSR, vol. 4, pp. 9-21, 1989.

[16] Laver, J., “The Phonetic Description of Voice Quality”, Cambridge University Press, Cambridge. 1980.

[17] K. Sri Rama Murthy and B. Yegnanarayana, “Epoch Extraction from Speech Signals,” IEEE Trans. Audio, Speech Lang. Process.,vol. 16, no. 8, pp. 1602-1613, Nov. 2008.

[18] K. Sri Rama Murty and B. Yegnanarayana and Anand Joseph M., “Characterization of Glottal Activity from Speech Signals,” *IEEE signal processing letters,* vol. 16, no. 6, June 2009.

[19] B. Yegnanarayana, C. R. d’Alessandro, and V. Darsinos,“An iterative algorithm for decomposition of speech signals into periodic and aperiodic components” IEEE Trans. Speech Audio Processing, vol. 6, pp.1-11, Feb. 1998.

[20] Anand Joseph M., Guruprasad S. and Yegnanarayana B., “Extracting Formants from Short Segments of Speech using Group Delay Functions”,Proc. Int. Conf. Spoken Language Processing (INTERSPEECH) 2006, pp. 1009-1012, Pittsburgh PA, USA, 17-21 Sept. 2006.

[21] G.Sheshadri and B.Yegnanarayana, “Perceived loudness of speech based on the characteristics of excitation source”, Journal of Acoustical Society of America, Vol. 126, No.4, pp. 2061-2071, Oct 2009.