Significance of incorporating excitation source parameters for improved emotion recognition from speech and electroglottographic signals

D. Pravena[1] · **D. Govind**[1]

Received: 20 April 2017 / Accepted: 29 July 2017 / Published online: 17 August 2017

Abstract The work presented in this paper explores the effectiveness of incorporating the excitation source parameters such as strength of excitation and instantaneous fundamental frequency (F_0) for emotion recognition task from speech and electroglottographic (EGG) signals. The strength of excitation (SoE) is an important parameter indicating the pressure with which glottis closes at the glottal closure instants (GCIs). The SoE is computed by the popular zero frequency filtering (ZFF) method which accurately estimates the glottal signal characteristics by attenuating or removing the high frequency vocaltract interactions in speech. The arbitrary impulse sequence, obtained from the estimated GCIs, is used to derive the instantaneous F_0. The SoE and the instantaneous F_0 parameters are combined with the conventional mel frequency cepstral coefficients (MFCC) to improve the recognition rates of distinct emotions (Anger, Happy and Sad) using Gaussian mixture models as classifier. The performances of the proposed combination of SoE and instantaneous F_0 and their dynamic features with MFCC coefficients are compared with the emotion utterances (4 emotions and neutral) from classical German full blown emotion speech database (EmoDb) having simultaneous speech and EGG signals and Surrey Audio Visual Expressed Emotion database (3 emotions and neutral) for both speaker dependent and speaker independent emotion recognition scenarios. To reinforce the effectiveness of the proposed features and for better statistical consistency of the emotion analysis, a fairly large emotion speech database of 220 utterances per emotion in Tamil language with simultaneous EGG recordings, is used in addition to EmoDb. The effectiveness of SoE and instantaneous F_0 in characterizing different emotions is also confirmed by the improved emotion recognition performance in Tamil speech-EGG emotion database.

Keywords Excitation source parameters · Emotion recognition · Strength of excitation · Zero frequency filtering

1 Introduction

The emotive content in the speech signal collected over a microphone provides extra linguistic information which glimpses on psychologic state of the speaker. By automatically extracting emotions from the spoken waveforms, many researchers have been exploring many applications such as virtual agents to automatically determine attitudes of the speakers and social behaviors Creed and Beal (2005); Ringeval et al. (2013); Cerezo (2007). For instance, Ringeval et al. (2013), developed a database based on remote collaborative and affective interactions which has annotations both in the affective and social dimensions. These annotations enable to automatically predict the social behavior of speakers from their conversations. The two stages involved in the development of the emotion recognition system are : (1) Analysis of the emotion dependent parameters and (2) modeling of these emotion dependent parameters for effective emotion recognition Ayadi et al. (2011). Compared to other parameters of the speech, the emotions are characterized by fine level variations (subsegmental) or variations

✉ D. Govind
d_govind@cb.amrita.edu

D. Pravena
d_pravena@cb.amrita.edu

[1] Center for Computational Engineering and Networking (CEN), Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Amrita University, Coimbatore, India

which are spanned over longer segments (at the suprasegmental levels) Govind et al. (2011); Erickson (2005); Ayadi et al. (2011); Govind and Prasanna (2013); Schroder (2009). There were many studies in the literature which demonstrate the effectiveness of subsegmental features such as parameters related to excitation features in characterizing different emotions Fairbanks and Hoaglin (1939); Cahn (1989); Cabral and Oliveira (2006); Whiteside (1998); Prasanna and Govind (2010); Govind et al. (2011). Therefore, the work presented in this paper explores different emotion dependent excitation parameters and proposes effective ways to incorporate these features for the improved emotion recognition.

Depending on sound to be spoken, the human speech is produced by the vibration of vocalcords or glottis and subsequent vocaltract interactions and most commonly known as the source-filter interactions Fant (1960). The speech systems developed for various speech processing tasks exploit the features representing the information at segmental Reynolds and Rose (1995), sub-segmental Pradhan and Prasanna (2013); Cabral and Oliveira (2006); Pati and Prasanna (2011); Prasanna and Yegnanarayana (2004) and supra-segmental levels Rao and Yegnanarayana (2006); Prasanna et al. (2010); Adiga and Prasanna (2013) in speech signals. Since the sub-segmental features are extracted within the pitch period of the speech signals, the features predominantly represent source component of speech Pati and Prasanna (2011); Govind and Prasanna (2013). In the absence of ground truth glottal flow signals, in most of the works, the sub-segmental features are estimated from linear prediction (LP) residual which is considered as the close approximation of the glottal flow derivative which is in turn obtained by the LP analysis of speech. The ground truth glottal flow derivatives can be measured with the help of electroglottograph by directly placing sensors around the glottis area and signal acquired in such a way is termed as electroglottographic (EGG) signals. The ground truth glottal parameter values can be directly computed from EGG signals. For instance, one of the important source parameters such as glottal closure instants (GCIs), can be computed by measuring the dominant discontinuity in the differenced EGG signal. In this way, the accuracy and precision of the glottal source parameters measured from speech can be compared with the ground truth parameter values computed from EGG Adiga and Prasanna (2008); Yegnanarayana and Murty (2009). With the availability of the EGG recordings for various emotions, the present work focusses on analyzing robust emotion dependent source parameters which characterize various emotions in EGG and speech.

The instantaneous pitch, jitter, shimmer and glottal flow parameters such as open quotient, speed quotient and return quotient etc. are categorized as the excitation source related features Cabral and Oliveira (2006); Whiteside (1998); Prasanna and Govind (2010). The instantaneous pitch period is computed as the interval between successive glottal closure instants in speech Yegnanarayana and Murty (2009). To compute instantaneous pitch period, the glottal closure instants in the speech have to be accurately estimated. The product of reciprocal of instantaneous pitch period and sampling frequency of the signal gives the instantaneous pitch (F_0). There are many works reported in the literature which discuss on the variation of pitch contours with emotions Prasanna and Govind (2010); Bulut and Narayanan (2008); Cabral and Oliveira (2006); Haq and Jackson (2010). Due to rapid and uncontrolled variations in instantaneous pitch in emotive speech, there are issues in the accurate estimation of GCIs which in turn affect the estimation of instantaneous pitch contours Govind and Prasanna (2012, 2013). The emotion dependent nature of jitter, an excitation parameter computed as the average difference in the instantaneous pitch values across successive pitch cycles, is studied by Whiteside (1998). Along with jitter, shimmer which provides average variation of intensity across successive pitch cycles, is also reported to vary consistently with various emotions. Cabral demonstrates the effectiveness of modifying instantaneous pitch contour, jitter and shimmer for neutral to emotion speech conversion in Cabral and Oliveira (2006). Cabral et al. also reports that the glottal flow parameters derived the LP residual namely open quotient, speech quotient and return coefficient also found to vary according to various emotions. In the present work, the effectiveness of strength of excitation (SoE), one of the important excitation source parameters characterizing the glottal pulse, in discriminating emotions is studied.

The SoE represents the suction pressure during the abrupt instants of glottal closure in voiced speech Murty and Yegnanarayana (2009). The SoE is computed by measuring the energy of glottal flow derivative at the GCIs. Since LP residual is considered as approximation of glottal flow derivative, the average energy of the residual samples around the GCIs are computed as SoE Rao and Yegnanarayana (2003). The effect of residual samples around the GCIs on the perceptual quality in the task of manipulation of prosodic parameters of speech is demonstrated by Rao et al. in Rao and Yegnanarayana (2006). These perceptually relevant residual samples clearly indicate the significance of preserving residual energy containing SoE for better perceptual quality for synthesis applications Rao and Yegnanarayana (2006); Govind and Joy (2016). Murty et al. proposed an accurate method for estimating SoE parameters by zero frequency filtering of speech signals Murty and Yegnanarayana (2009). To reduce the effect of high frequency vocaltract interactions and emphasize the impulse like discontinuities at the GCIs, the speech signal is passed through a cascade of two resonators whose resonance center frequency is located at 0 Hz. To obtain the low frequency variations at the source level, the local mean subtraction is performed on the output of the zero frequency resonator. The positive zero crossings of the resulting mean subtracted zero frequency filtered signal (ZFFS) are estimated as GCIs and the positive slope at the GCIs are hypothesized as the corresponding strength of excitation parameters. The estimated GCIs are observed to coincide accurately with the reference GCIs obtained from the differenced EGG signal and the SoE parameters estimated from zero frequency filtered signal are linearly proportional to the reference parameter values measured from the EGG Murty and Yegnanarayana (2009).

Prasanna et al. reported the variations in instantaneous F_0 according to different emotions in speech and EGG Prasanna and Govind (2010). Improvement in the emotion recognition performance is achieved using source parameters such as instantaneous F_0, SoE and energy of excitation Kadiri et al. (2015). As the emotion recognition system developed by modeling the source parameters deviations for each emotion from the neutral speech of the given speakers, the system requires the availability of neutral utterances from each speaker before testing. However, the purpose of the study was to demonstrate the significant emotion information carried by excitation source parameters. Motivated from these works, the present work studies the independent and combined effect of concatenating the SoE parameters and instantaneous F_0 with the spectral parameters represented by mel frequency cepstral coefficients (MFCC). As a part of the paper, the effect of dynamics of the source parameters in discriminating different emotions are also studied. For testing the generality of the methods proposed in the work, a comprehensive simulated full blown emotion database is developed with simultaneous speech and EGG signals for distinct emotions in Tamil language. To ensure the perceptual quality in the recordings, the methods adopted for the proper elicitation of the emotions by the subjects and comparative studies are conducted for the emotive quality with the classical simulated emotion databases. The rest of the paper is organized as follows: The details of the simulated full blown emotion database is explained in the Sect. 2. The significance of strength of excitation parameter characterizing different emotions is given in Sect. 3. Development of proposed emotion recognition system by merging strength of excitation and its dynamic coefficients is explained in Sect. 4. Finally Sect. 5 summarizes the work presented in the paper with scope for future works.

2 Speech and EGG emotion databases used for the experiments

As the first stage for the development of the emotion recognition system is the emotion analysis stage, the emotion data using which analysis is performed for the emotion dependent parameters are going to be crucial. For the present work, the classical German simulated speech emotion database Burkhardt et al. (2005) and recently developed multilingual simulated speech emotion database Pravena and Govind (2017) are used for the experimental analysis and performance evaluation.

2.1 German speech emotion database [EmoDb; Burkhardt et al. (2005)]

German EmoDb is considered as the classical database for the analysis of emotions in speech. EmoDb consists of seven emotions spoken by ten professional speakers. In EmoDb, speakers elicited emotions on the neutral German utterances during the recording stage Burkhardt et al. (2005). Each emotion utterance in the database has simultaneous speech and EGG recordings. Four emotions (Happy, Anger, Boredom and Fear) apart from neutral utterances are selected from EmoDb for the work presented in this paper. Based on perceptual distinction, the four emotions of EmoDb are selected from seven emotions for the excitation source analysis. EmoDb has on an average 100 emotion utterances per each emotion category.

2.2 Multilingual simulated speech emotion database for the indian context [Pravena and Govind (2017)]

Multilingual speech emotion database has three perceptually distinct emotions (Anger, Happy and Sad) apart from neutral utterances for three languages such as Tamil, Malayalam and Indian English. With the availability of the EGG recordings along with speech, the database is large and particularly developed for excitation source analysis in the Indian context. Each language has nearly ten amateur speakers and 220 emotionally biased (utterances selected from emotion dependent contexts) utterances recorded in two sessions. Based on the emotion recognition experiments conducted in Pravena and Govind (2017), the speakers tend to elicit emotions in a better way when they simulated emotionally biased utterances as compared to emotionally neutral utterances during the recording stage. The speech and EGG utterances selected from one session of the Tamil language are used for the emotion analysis studies presented in this paper.

2.3 Surrey audio visual expressed emotion (SAVEE) database [Haq and Jackson (2009, 2010)]

Surrey audio visual expressed emotion (SAVEE) database consists of 120 utterances spoken by four professional actors in six emotions (Anger, Happy, Sad, Disgust, Fear, Surprise) and Neutral. 30 neutral utterances and 15 emotion utterances for each emotion category constitute a total of 120 utterances (30 + 15 × 6) per speaker. 15 neutral utterances from TIMIT database, three common

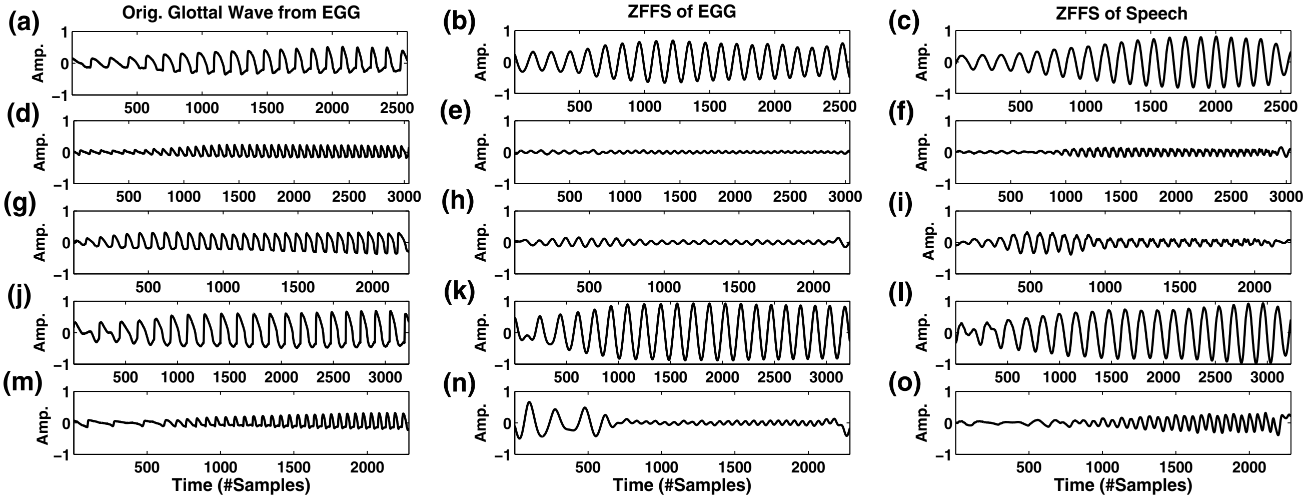

Fig. 1 Comparison of original glottal waveform obtained from EGG, ZFFS obtained from EGG and ZFFS obtained from speech for neutral (**a**–**c**), anger (**d**–**f**), happy (**h**, **i**), boredom (**j**–**l**) and fear (**m**–**o**),respectively.

phonetically balanced utterances, two emotion specific utterances for each emotion category are recorded by each speaker to make a total of neutral utterances per neutral and 15 utterances per emotion category. Even though the size of the SAVEE database is similar to German EmoDb, the database contains video recordings corresponding to each emotive audio utterance recording.

3 Significance of excitation source features in characterizing emotions

The instantaneous pitch and strength of excitation are the important source features considered for the present work. Both the source features are extracted by processing the zero frequency filtered signal (ZFFS) obtained by the zero frequency filtering (ZFF) of speech. As the high frequency vocaltract responses are attenuated in ZFF method by filtering the speech signal though a zero frequency resonator, the predominant characteristics of the glottal signal are well preserved in the resulting ZFFS. A comparison of glottal waveform characteristics with original glottal waveform obtained from EGG, ZFFS computed by the ZFF of EGG and ZFFS obtained from speech is provided in Fig. 1 for different emotions. The original glottal wave characteristics of neutral emotion plotted in Fig. 1a show considerable variations in terms of the number of glottal cycles, amplitude and shape as compared to the other emotions plotted in Fig. 1d, g, j and m.

3.1 Observations on glottal waveform characteristics due to emotions

The following are the observations on the glottal waveform characteristics from Fig. 1, there is a strong dependency of various emotions on the number of glottal cycles present in each of waveform segments which in turn represents variations in the instantaneous F_0. The emotion dependent variations on the SoE features are also observed from Fig. 1. For instance, Compared to neutral glottal waveform segment plotted in Fig. 1a has reduced number of glottal cycles as compared to that of the anger glottal waveform segment plotted in Fig. 1d. However, in the case of the amplitudes of the glottal pulses for neutral and anger emotions, the trend is the reverse. For instance, the average amplitude of the neutral glottal cycles is higher (from Fig. 1a–c) as compared to that of the anger glottal waveforms (from Fig. 1d–f). In all the subplots of Fig. 1, the ZFFS segments of EGG and speech show similar glottal characteristics as indicated in the original glottal segments. Form the speech production view point, reasons for the lower glottal amplitudes in anger emotions are due to the vibration of glottis with reduced pressure due to increase in the pitch. As the periodic glottal vibrations with lower fundamental frequencies always increase glottal contact areas between the vocal folds, the amplitude of the corresponding glottal cycles are higher for neutral waveform segments compared to other emotive waveform segments having higher F_0 values. The aforesaid observations are reinforced when average F_0 and SoE parameters of neutral and boredom emotion segments are compared from

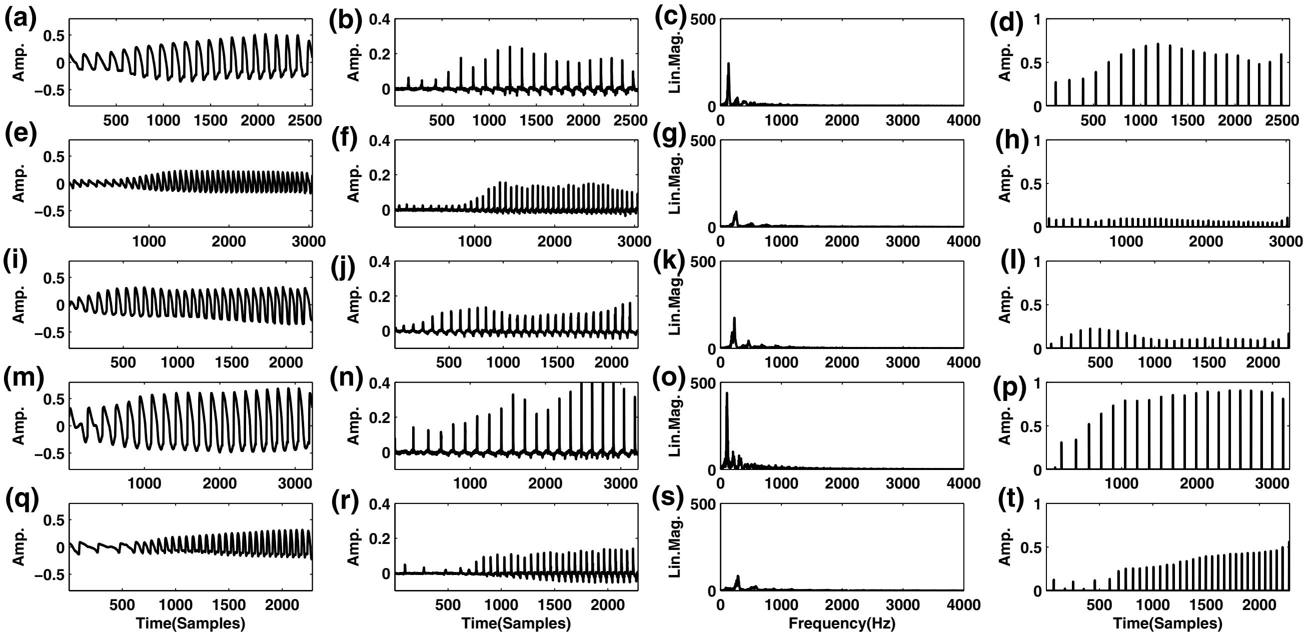

Fig. 2 Variations in source characteristics of EGG signals in different emotions. The EGG signal waveform, differenced EGG waveform, STFT of EGG and Strength of Excitation of neutral (**a**–**d**), anger (**e**–**h**), happy (**i**–**l**), boredom(**m**–**p**) and fear (**q**–**t**) utterances. The waveform segments are selected from the same syllable of the same sentence spoken by the same speaker of German EmoDb database

Fig. 1 where in lower F_0 and higher SoE values are obtained for both the cases.

The subplots in second and third columns of Fig. 1 show the corresponding ZFFS segments obtained from EGG and speech, respectively. By comparing the original glottal waveforms obtained from EGG, the corresponding ZFFS segments estimated from speech show corresponding variations. For instance, the variations of the ZFFS segments in Fig. 1c for neutral emotion is proportional to the corresponding original EGG segments in terms of the number of pitch cycles and energy levels. The trend in the emotive glottal characteristics observed for EGG segments are observed in the case of ZFFS segments from speech also by comparing different emotions. The ZFFS estimated from EGG also shows the exactly same variations as in the original EGG signals which further reinforces the effectiveness of ZFFS in preserving the glottal waveform characteristics. Having analyzed the properties preserving glottal characteristics of EGG, ZFFS derived from speech is used for the source analysis across different emotions.

3.2 Analysis of excitation parameters across different emotions [Prasanna and Govind (2010)]

Based on the visual inspection of glottal waveform characteristics across different emotions and as plotted in Fig. 1, source features of interest in study are instantaneous pitch and strength of excitation parameters (SoE). Figure 2 plots the variation of SoE across different emotions. As SoE is proportional to the energy with which glottis vibrates and is indicated as the dominant peak at the GCI in the differenced EGG (subplots in the 2nd column of Fig. 2). As given in the 3rd column of subplots in Fig. 2, the peak magnitude of the STFT provides the average SoE magnitude of the short time segment of EGG. Instantaneous SoE at every GCI is measured as the slope of the zero frequency filtered signal obtained by the ZFF of EGG and is plotted as the 4th column of subplots in Fig. 1. The average trend in the variations of SoE follow the same way as discussed earlier for different emotions. For instance, by comparing Fig. 2d and Fig. 2h, anger emotion has lower SoE as compared to that of the neutral due to the lower contact area of the glottis at the closure because of the higher fundamental frequency. Similarly, higher values of SoE indicate the higher suction pressure at the time of glottal closure as compared to other emotions due to lower pitch. The variations in the average pitch and the average SoE corresponding for the syllable segment of each emotion can be observed from the magnitude STFT plot. For instance, the subplots of the 3rd column of the Fig. 2, the average pitch is shown as the frequency corresponding to the largest magnitude. Analysis of STFT plot also reinforces the observations from the instantaneous variations of the SoE represented by the 2nd and 4th columns of Fig. 2 in an average sense. Also, the lowest pitch and highest SoE values are obtained for boredom emotions (from the plot of Fig. 2o and p.

Table 1 The gross level variations of instantaneous pitch and strength of excitation parameters estimated from speech

Emotions	EmoDb				Tamil DB			
	Mean pitch		SoE		Mean pitch		SoE	
	EGG	SPEECH	EGG	SPEECH	EGG	SPEECH	EGG	SPEECH
Neutral	181.74	189.34	0.60	0.48	187.39	195.83	0.41	0.47
Anger	272.29	291.12	0.49	0.39	211.68	226.92	0.34	0.42
Happy	238.42	256.52	0.54	0.41	205.71	229.13	0.32	0.40
Boredom	168.87	178.63	0.62	0.54	–	–	–	–
Fear	193.26	223.01	0.31	0.43	–	–	–	–
Sad	–	–	–	–	174.94	184.45	0.46	0.49

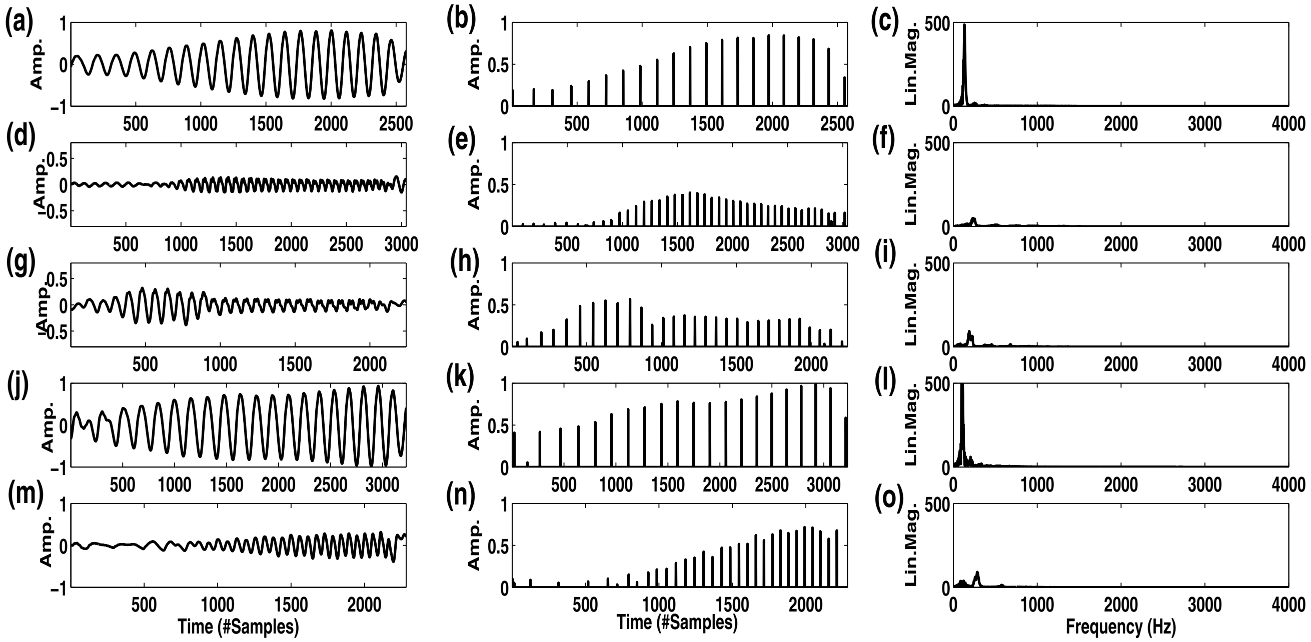

Fig. 3 Variations in source characteristics of zero frequency filtered signal (ZFFS) estimated from speech (approximated glottal waveform equivalent estimated from speech) in different emotions. The ZFFS segment, estimated SoE parameter values and linear magnitude spectrum of ZFFS segment of neutral (**a–c**), anger (**d–f**), happy (**g–i**), boredom(**j–l**) and fear (**m–o**) utterances. The speech waveform segments are selected from the same syllable of the same sentence spoken by the same speaker of German EmoDb database

Figure 3 shows the plots of SoE and magnitude of STFT showing the average pitch obtained from ZFFS of speech. Here also the variations in SoE computed from ZFFS show variations in accordance with original SoE represented by the differenced EGG across different emotions. Based on the analysis of subplots in the second column of the Fig. 3, the lowest SoE is obtained for anger emotion (Fig. 3e. The lowest SoE for anger is also confirmed by observing the magnitude of STFT plotted in subplot (f) of Fig. 3. By observing the STFT magnitude plots of ZFFS in Fig. 3, the trend in the variations of pitch obtained is also fall in accordance with the values computed from the original EGG.

The gross level quantitative comparisons of SoE and average instantaneous pitch parameters for EmoDb and Tamil emotion database are given in Table 1. The pitch and SoE parameters are estimated from ZFFS of speech and EGG. The gross level variations of SoE and instantaneous pitch are computed by taking the average across all utterances for different emotion categories. The analysis on average variations of the parameters across different emotions also reinforces the noted observations from the Figs. 2 and 3. For instance, reduced average SoE and higher average pitch values are consistently obtained for anger emotions in both the databases. Similarly, boredom emotion of EmoDb and sad emotions of Tamil emotion database provided highest average SoE and lowest mean pitch, respectively. After boredom emotion in EmoDb and sad emotion in Tamil database , the highest mean SoE values are obtained for neutral emotions of both the databases. Higher average pitch values of happy emotions in EmoDb and Tamil database which is comparable to that of anger, the average SoE obtained is also lower. In the Table 1, the SoE and pitch parameters computed from EGG also observed to vary proportionally with that of speech. However, the parameter values obtained from speech and EGG show differences in their values computed. The reason is the presence of well defined impulse like discontinuities correspond to GCIs in differenced EGG when compared to speech where the discontinuities are smoothed due to the vocaltract interactions. However, the difference is less for the average parameter values of speech and EGG in the case of neutral utterances as compared to the rest of the emotions except boredom and sad emotion utterances. Due to the rapid variations in pitch for anger, happy and fear emotions, there are issues in the estimation of window length which in turn affect the accuracy of the ZFFS estimated. This issue of inaccurate ZFFS estimation from emotion utterances such as anger, happy and fear is reported and addressed by Kadiri et al. in Kadiri and Yegananarayana (2015). Due to the lack of rapid pitch variations in neutral, boredom and sad emotions, the ZFFS estimated and parameter values computed from speech and EGG show only slight differences. However, the trend observed about the variations in SoE and pitch values show consistent variations across different emotions. By analyzing the average parameter values obtained for happy and fear emotions from the Table 1, show emotion dependent variations. Hence, these emotion dependent variations have to be incorporated along with the commonly used speech features to improve the emotion recognition rates of the classical systems.

4 Merging excitation source parameters for improved emotion recognition in speech and electro-glottogram signals

To understand the effect of excitation source parameters in emotion recognition, the instantaneous F_0 and SoE parameters are explicitly combined with the classical mel frequency cepstral coefficients (MFCC) features (39 dimensional MFCC coefficients with velocity, Δ and acceleration, $\Delta\,\Delta$ coefficients).[1] The computed source features are concatenated with each frame of MFCC coefficients. As the SoE parameters are discrete in nature (as SoEs are computed only at corresponding GCIs), the linear interpolation is performed between successive GCIs. In this way the SoE parameters are made available continuously for every time instant of the speech or EGG waveform. The average of the SoE values are then computed corresponding to each frame of the signals and concatenated with the MFCC features extracted from the corresponding frame. To match dynamic range of the MFCC coefficients, the logarithm of the average SoE values obtained for each frame is computed before concatenating with the corresponding MFCC frames. In the similar manner, the logarithm of mean instantaneous pitch computed for each frame is combined with MFCC features. The average pitch values computed for each frame that are falling outside the human pitch range(70–300 Hz) are replaced with some constant pitch value. The pitch level threshold is set for removing the spurious instantaneous F_0 values which are possibly introduced due to falsified GCI estimation from different emotions.

Figure 4 shows the emotion recognition performance obtained for different combinations of instantaneous F_0 and SoE features along with MFCC coefficients. The bar plot is generated for both speech and EGG obtained from EmoDb and Tamil database with emotionally biased utterances. The bar plot shows the gross level recognition rate of different emotion classes available in EmoDb and Tamil emotional databases. The first level analysis of the results shows that the concatenation of instantaneous pitch, SoE and their velocity and acceleration (Δ and $\Delta\,\Delta$) parameters along state of the art MFCC parameters (with Δ and $\Delta\,\Delta$) significantly improve the gross level emotion recognition performance in both speech and EGG cases. By observing the Fig. 4, there are variations in the recognition rates when instantaneous pitch and SoE parameters are independently concatenated with 39 MFCC features. For instance, in the case of EmoDb, $MFCC + F_0$ and $MFCC + SoE$ showed gross level emotion recognition accuracy of 74.14 and 75.0%, respectively of five emotion classes. Even though the differences in the recognition performances between the two cases are minimal or equal, significant differences in the recognition rates of individual emotion classes are observed (from Table 2). The complimentary emotion information carried by instantaneous pitch and SoE is the reason for the difference in emotion wise accuracies in both the cases. As a result of this improved gross level emotion recognition rate is obtained for MFCC+F_0+SoE feature combination (third set of bars in Fig. 4a). The same trend can also be observed for EGG available in EmoDb from Fig. 4b. In the case of EGG also, the combination of MFCC, F_0 and SoE provided improved recognition rate than the their individual combinations with MFCC. Even though the dynamic features (Δ and $\Delta\,\Delta$) of SoE show improved gross level emotion recognition rate, the improvement is evident mostly in the neutral class. For instance, from the Fig. 4b, an absolute improvement in the recognition rates for $MFCC + SoE$ (70.69%) and $MFCC + SoE$+Δ SoE+$\Delta\,\Delta$ SoE (74.14%) can be observed for the features obtained from EGG signals available in

[1] MFCC term used throughout this paper denotes 39 MFCC coefficients having 13 MFCC along with 13 velocity (Δ) and 13 acceleration ($\Delta\,\Delta$) coefficients.

Table 2 The emotion class wise performance obtained for speech and EGG signals by merging SoE and F_0 parameters with 39 MFCC coefficients for EmoDb and Tamil emotion speech database

Features	SPEECH										
	EmoDb						Tamil DB				
	Overall accuracy (%)	Class wise accuracy (%)					Overall accuracy (%)	Class wise accuracy (%)			
		Neutral	Anger	Happy	Boredom	Fear		Neutral	Anger	Happy	Sad
MFCC	73.28	85.0	100	46.2	72.7	60.0	83.52	90.7	69.5	82.5	91.4
MFCC+F_0	74.14	85.0	96.4	53.8	63.6	70.0	85.23	93.2	71.1	85.9	90.7
MFCC+SoE	75.00	85.0	100	50.0	68.2	70.0	82.27	90.5	68.9	81.4	88.4
MFCC+F_0+SoE	75.86	85.0	92.9	65.4	63.6	70.0	84.32	90.7	71.6	86.6	88.4
MFCC+F_0+Δ F_0+Δ Δ F_0	78.45	85.0	96.4	61.5	72.7	75.0	86.14	94.3	70.9	89.1	90.2
MFCC+SoE+ΔSoE+Δ ΔSoE	74.14	85.0	96.4	57.7	63.6	65.0	84.43	90.5	70.9	85.7	90.7
MFCC+F_0+Δ F_0+Δ Δ F_0 +MFCC+SOE+ΔSOE+Δ Δ SOE	78.45	90.0	89.3	69.2	72.7	70.0	85.23	90.9	72.7	87.7	89.5
EGG											
MFCC	67.24	60.0	75.0	61.5	68.2	70.0	74.60	88.0	60.0	75.5	74.1
MFCC+F_0	65.52	65.0	75.0	61.5	63.6	60.0	76.88	89.8	59.8	78.6	79.3
MFCC+SoE	70.69	70.0	85.7	57.7	59.1	80.0	76.36	88.4	62.7	77.5	76.8
MFCC+F_0+SoE	72.41	70.0	82.1	69.2	68.2	70.0	77.44	89.5	62.7	77.7	79.8
MFCC+F_0+Δ F_0+Δ Δ F_0	63.79	60.0	71.4	53.8	68.2	65.0	77.67	91.4	62.7	77.7	78.9
MFCC+SoE+ΔSoE+Δ ΔSoE	74.14	80.0	82.1	65.4	63.6	80.0	77.50	88.4	65.2	78.4	78.0
MFCC+F_0+Δ F_0+Δ Δ F_0 + +SoE+ΔSoE+Δ ΔSoE	74.14	75.0	82.1	69.2	68.2	75.0	79.38	90.0	65.2	81.4	80.9

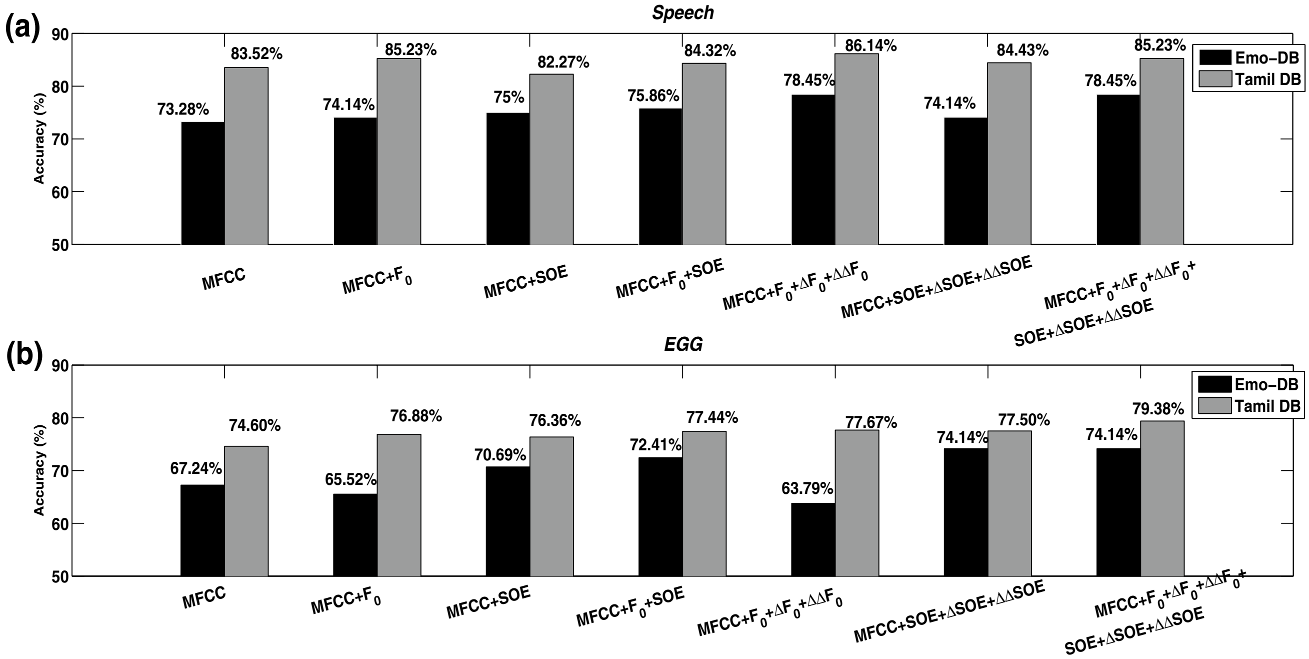

Fig. 4 The emotion recognition performance obtained for combinations of SoE and pitch at various levels with MFCC features using GMMs in **a** speech and **b** EGG database obtained from EmoDb and Tamil emotion databases

Table 3 Average speaker independent emotion recognition for the proposed features on German EmoDb and SAVEE databases

Features	EmoDb						SAVEE				
	SPEECH										
	Overall accuracy (%)	Class wise accuracy (%)					Overall accuracy (%)	Class wise accuracy (%)			
		Neutral	Anger	Happy	Boredom	Fear		Neutral	Anger	Happy	Sad
MFCC	72.14	72	90.1	57.1	83.2	58.4	58.22	73.3	69.8	63.3	26.6
MFCC+F0	73.45	72.2	91.5	53.6	86.1	61.1	61.49	78.3	70	71.8	27.0
MFCC+SOE	74.45	74.12	93.8	64.1	86.3	61.8	62.05	75.0	71.7	70	36.7
MFCC+F0+SOE	76.86	77.9	91.4	62.3	87.0	63.4	63.00	81.7	75.0	71.7	34.3
MFCC+F0+ΔF0+ΔΔF0	72.39	67.8	92.2	57	84.8	58	65.83	88.3	68.3	75.0	38.3
MFCC+SOE+ΔSOE+ΔΔSOE	75.51	76.1	90.7	53.6	87.7	66.7	66.25	85	76.7	70.0	85.0
MFCC+F0+ΔF0+ΔΔF0 + +SOE+ΔSOE+ΔΔSOE	77.51	74.04	93.2	66.6	86.8	64.4	70.43	83.3	80.0	76.1	45.1

EmoDb. Comparing the emotion wise performances corresponding to the features $MFCC + SoE$ and $MFCC + SoE+\Delta$ $SoE+\Delta\,\Delta\, SoE$ from the Table 2, the absolute improvement in the overall accuracy due to the overshoot of recognition rate of neutral class. A similar trend can be observed for dynamic features of F_0 obtained from speech signals of Tamil emotion database.

Tamil emotion speech database also provided similar trends when the instantaneous F_0 and SoE features along with their dynamics are combined with MFCC features for speech as well as EGG. As obtained for EmoDb, the effect of dynamic features of F_0 and the derivatives are observed to be significant in the case of Tamil emotion database also. However the effect of SoE in improving the gross level emotion classification rate for Tamil speech is not as evident as observed in the case of EmoDb. Given the equal recognition rates obtained for neutral class, the reason for the degradation is due to inconsistencies in the estimation of SoE parameters from different emotion classes. The SoE features obtained from EGG data of Tamil emotion database are found to show improved performance in gross level emotion recognition rates. By observing Fig. 4 and Table 2, the $MFCC + F_0$ and $MFCC + F_0 + \Delta F_0 + \Delta\Delta F_0 + SOE + \Delta SOE + \Delta\Delta SOE$, provided the same speech emotion recognition rate. However, performance variations for the individual emotion classes are observed by the independent class wise analysis of Table 2. The same average speech emotion recognition rate with varied class level accuracies shows the evidence of complimentary emotion information present in instantaneous F_0 and SoE features. Even though, improvement in the gross level recognition rate is obtained for $MFCC + F_0 + \Delta F_0 + \Delta\Delta F_0$, class wise performance Table shows, improvement is significantly confined to the neutral class alone with minimal variations in the other emotion classes. This trend can be observed for the case of available Tamil emotion EGG data also. In summary, by analyzing the emotion recognition rates at the gross level (given in Fig. 4) and individual emotion class wise (given in Table 2), the instantaneous F_0 and SoE features carry information related to emotion classes and hence improved recognition performance.

4.1 Speaker independent speech emotion recognition performance evaluation

In the speaker independent emotion recognition experiment, recognition performances are obtained for the utterances from the speakers who were unseen during the training. Since, the EmoDb and SAVEE databases are similar in terms of the number of speech emotion utterances, the speaker independent recognition rates of both the databases are compared. Also, speaker independent emotion recognition rates for the proposed features are validated by building GMMs with 256 Gaussian mixture components. Table 3 shows the speaker independent speech emotion recognition performance comparison for the proposed features on EmoDb and SAVEE databases. For EmoDb, GMMs with 256 mixure components are built by leaving two speakers and recognition rates are obtained for each of the training and recognition. A total of five such recognition trails are performed. The speaker independent emotion recognition rate obtained for each emotion category for EmoDb in Table 3 is the average value computed across five recognition trails. The emotion recognition rates provided in the Table 3 for SAVEE database are the average values computed across the four experimental trails with each of the four unseen speakers used for testing in each experimental trail. Equal number of utterances (15 utterances) from three speakers are used for building GMM models for each emotion category in every experimental trail. Out of the six emotions in the SAVEE database, three distinct emotions (Anger, Happy and Sadness) and neutral utterances are considered for the present work.

As compared to the speaker dependent emotion recognition obtained for EmoDb in Table 2, the overall trend in the emotion recognition rates remains the same in the speaker independent case also. Moreover, the speaker independent emotion recognition on SAVEE database also showed improved performance for the features obtained by merging SoE and F_0 parameters with 39 MFCC features. As obtained for the speaker dependent case, merging of the dynamic features derived from SoE and F_0 parameters with MFCC+ Δ + Δ Δ coefficients provided significant improvements over EmoDb and SAVEE databases in the speaker independent scenarios as well.

5 Summary and conclusions

The primary objective of the present work is to study the significance of important excitation source features such as instantaneous F_0 and strength of excitation in emotion recognition from speech and electro-glottographic signals. The instantaneous F_0 and SoE parameters which are estimated using the ZFF method, are concatenated with conventional 39 MFCC features to generate new features for developing emotion recognition systems. Along with the state of the art German EmoDb and SAVEE databses, a large Tamil emotion speech and EGG database (part of multilingual simulated speech emotion database) is prepared for the performance evaluation. Unlike the German EmoDb and SAVEE databases, the emotionally biased utterances are used for recording the relatively large Tamil emotion speech and EGG database. Based on the comparative performance analysis on the EmoDb and Tamil simulated speech emotion databases, the proposed features obtained by combining SoE and instantaneous F_0 are confirmed to carry significant emotive information invariably of the language and the type of emotion elicitation used for both speech and EGG signals. Moreover, the proposed features obtained by merging SoE, instantaneous F_0 and corresponding dynamic features along with conventional 39 MFCC coefficients (with dynamic features) are observed to provide effective emotion discrimination in the speaker independent scenarios as well.

The contributions of the paper are as follows:

- New features are proposed by merging emotion dependent excitation source feature such as instantaneous F_0 and strength of excitation (SoE) along with the state of the art MFCC features for emotion recognition
- In the context of emotion recognition, the instantaneous F_0, SoE and their dynamic features are observed to carry complimentary emotion information and hence showed improvement in emotion recognition rates when combined with MFCC features

As the source features in emotion speech vary rapidly, there are issues discussed in the literature for the accurate estimation of instantaneous pitch and SoE Govind and Prasanna (2012). For the experimental study presented in the paper, the issues with accurate estimation of instantaneous pitch and SoE are ignored. As a result, during the performance analysis of recognition rates of various emotions, some inconsistencies are observed. These performance inconsistencies are much more evident for the case of SoE than instantaneous F_0. In an emotive speech utterance, the instantaneous pitch values computed for a given speaker can vary only within a fixed range (within F_{0Min} and F_{0Max}). Hence, the margin of error in the case of falsified F_0 estimations are limited and constrained to defined F_0 range. Since there exists no such limit for SoE values, margin of variations are higher as compared to pitch. Hence, the falsified estimation of SoE values affect the recognition performance adversely than pitch. The consistent recognition rates observed for neutral class in all the cases reinforce the presence of erroneous SoE values in the case of emotions. As the source features in the EGG are defined in better way than speech, the ambiguity in the emotion recognition rates are less as compared to the speech case. As a future work, the issues in the accurate estimation of source features on the emotion recognition performances have to be studied in detail. Based on these studies, methods have to be devised for the accurate estimation of source parameters from emotive speech utterances. As an immediate extension of the work presented in the paper, the effectiveness of the proposed features has to be studied on an extended Tamil emotion speech database having more emotionally biased utterances and speakers. Also, validity of the results obtained in the present work has to be verified with emotion speech-EGG database collected from different languages.

Acknowledgements Works carried out for this paper are funded by the completed DST-SERB project titled, “Analysis, Processing and Synthesis of Emotions in Speech (Ref No. SB/FTP/ETA-370/2012)”.

References

Adiga, N. & Prasanna, S. R. M. (2013). Significance of instants of significant excitation for source modeling. In *Proceedings of INTERSPEECH*.

Ayadi, M. E., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes and databases. *Pattern Recognition*, *44*, 572–587.

Bulut, M., & Narayanan, S. (2008). On the robustness of overall f0 only modifications to the perception of emotions in speech. *The Journal of the Acoustical Society of America*, *123*, 4547–4558.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlemeier, W., & Weiss, B. (2005). A database of German emotional speech. In *Proceedings of INTERSPEECH* (pp. 1517–1520).

Cabral, J. P., & Oliveira, L. C. (2006). Emo voice: A system to generate emotions in speech. in *Proceedings of the INTERSPEECH* (pp. 1798–1801).

Cahn, J. E. (1989). Generation of affect in synthesized speech. In *Proceedings of the American voice I/O society* (pp. 1–19).

Cerezo, E. & Baldassarri, S. (2007). Interactive agents for multimodal emotional user interaction. In *In Proceedings of the international conference on interfaces and hman computer interaction*.

Creed, C., & Beal, R. (2005). Using emotion simulation to influence user attitudes and behaviors. In *Proceedings of workshop on role of emotion in HCI*.

Erickson, D. (2005). Expressive speech: Production, perception and application to speech synthesis. *Acoustical Science and Technology*, *26*(4), 317–325.

Fairbanks, G., & Hoaglin, L. W. (1939). An experimental study of pitch characteristics of voice during the expression of emotion. *Speech Monographs*, *6*, 87–104.

Fant, G. (1960). *Acoustic theory of speech production*. s-Gravenhage: Moutan & Co.

Govind, D., & Joy, T. T. (2016). Improving the flexibility of dynamic prosody modification using instants of significant excitation. *International Journal of Circuits Systems and Signal Processing*, *35*(7), 2518–2543.

Govind D. & Prasanna, S. R. M. (2012). Epoch extraction from emotional speech. In *Proceedings of signal procesing & communications (SPCOM)* (pp. 1–5).

Govind, D., & Prasanna, S. R. M. (2013). Expressive speech synthesis: A review. *International Journal of Speech Technology*, *16*(2), 237–260.

Govind, D. , Prasanna, S. R. M., & Yegnanarayana B. (2011). Neutral to target emotion conversion using source and suprasegmental information. In *Proceedings of INTERSPEECH 2011*.

Haq, S., & Jackson, P. J. B. (2009). Speaker-dependent audio-visual emotion recognition. in *Proceedings of international conference on audio visual speech processing* (pp. 53–58).

Haq, S., & Jackson, P. J. B. (2010). Chapter 17: Multimodal emotion recognition. In W. Wang (Ed.), *Machine audition: Principles, algorithms and systems*. Hershey: IGI Global Press.

Kadiri, S. R., Gangamohan, P., & Yegnanarayana, B. (2015). Analysis of excitation source features of speech for emotion recognition. in *Proceedings of INTERSPEECH*

Kadiri, S. R. & Yeganarayana, B. (2015). Analysis of singing voice for epoch extraction using zero frequency filtering method," in *International conference on acoustics, speech and signal processing (ICASSP)*.

Murty, K. S. R., & Yegnanarayana, B. (2008). Epoch extraction from speech signals. *IEEE Transactions on Audio Speech and Language Processing*, *16*(8), 1602–1614.

Murty, K. S. R., & Yegnanarayana, B. (2009). Characterization of glottal activity from speech signals. *IEEE Signal Processing Letters*, *16*(6), 469–472.

Pati, D., & Prasanna, S. R. M. (2011). Subsegmental, segmental and suprasegmental processing of linear prediction residual for speaker information. *International Journal of Speech Technology*, *14*(1), 49–64.

Pradhan, G., & Prasanna, S. R. M. (2013). Speaker verification by vowel and nonvowel like segmentation. *IEEE Transactions on Audio Speech and Language Processing*, *21*(4), 854–867.

Prasanna, S. R. M. & Govind, D. (2010). Analysis of excitation source information in emotional speech," in *Proceedings of the INTERSPEECH* (pp. 781–784).

Prasanna, S. R. M., & Yegnanarayana, B. (2004). Extraction of pitch in adverse conditions. In *Proceedings of ICASSP, Montreal*.

Prasanna, S. R. M., Govind, D., Rao, K. S., & Yenanarayana, B. (2010). Fast prosody modification using instants of significant excitation. In *Proceedings of speech prosody*.

Pravena, D. & Govind D. (2017). Development of simulated emotion speech database for excitation source analysis," *International Journal of Speech Technology*. DOI:10.1007/s10772-017-9407-3.

Rao, K. S., & Yegnanarayana, B. (2006). Prosody modification using instants of significant excitation. *IEEE Transactions on Audio Speech and Language Processing*, *14*, 972–980.

Rao, K. S. & Yegnanarayana, B. Prosodic manipulation using instants of significant excitation. In *Proceedings of ICASSP* (pp. 528–531).

Reynolds, D., & Rose, C. (1995). Robust text independent speaker recognition using gaussian mixture speaker models. *IEEE Transactions on Audio Speech and Language Processing*, *3*(1), 72–83.

Ringeval, F., Sonderegger A., Sauer J., & Lalanne D. (2013). Introducing the recola multimodal corpus of remote collaborative and affective interactions, In *2nd international workshop on emotion representation, analysis and synthesis in continuous time and space (EmoSPACE), in Proceedings of IEEE Face & Gestures*.

Schroder, M. (2009). Expressive speech synthesis: Past, present and possible futures. *Affective information processing* (pp. 111–126). Berlin: Springer.

Whiteside, S. P. (1998). Simulated emotions: An acoustic study of voice and perturbation measures. *Proceedings of the ICSLP*, Sydney (pp. 699–703).

Yegnanarayana, B., & Murty, K. S. R. (2009). Event-based instantaneous fundamental frequency estimation from speech signals. *IEEE Transactions on Audio Speech and Language Processing*, *17*(4), 614–625.