A Flexible Analysis Synthesis Tool (FAST) for studying the characteristic features of emotion in speech

P. Gangamohan[1] *V. K. Mittal*[2] *and B. Yegnanarayana*[3]

International Institute of Information Technology, Hyderabad

[1]gangamohan.p@students.iiit.ac.in [2]vinay.mittal@iiit.ac.in, [3]yegna@iiit.ac.in

Abstract

This paper aims to understand the components of speech that contribute to emotion characteristics in speech. Four components of speech (vocal tract, excitation, duration and intonation) are considered in this study. A Flexible Analysis Synthesis Tool (FAST) is developed to modify the features of an utterance from neutral to emotion or from emotion to neutral. The key ideas used in this work are the dynamic time warping algorithm for alignment of two utterances and a flexible prosody manipulation for incorporating the desired features. The tool is used for conversion of neutral to emotion speech. Subjective evaluation is performed based on listening tests. The tool has potential to convert neutral to emotion speech and vice-versa, which can lead to understanding the significance of various components contributing to emotional content in speech.

Index Terms: Emotion analysis, emotion synthesis, emotion conversion, dynamic time warping, zero frequency filtering, prosody modification.

1. Introduction

The objective of this study is to understand the components of speech that contribute to emotion characteristics in speech. For this study the data is collected from different speakers, each speaker uttering a given sentence in normal (neutral) mode and in emotional mode [1]. The utterances are aligned using dynamic time warping (DTW) algorithm. The DTW path is used to determine the corresponding frames between neutral and emotional utterances [2]. The features of speech that can be studied are: vocal tract features represented by linear prediction coefficients [3], excitation features represented by the linear prediction residual [4], duration features represented by the warping path and intonation features represented by the pitch contour [5].

It is possible to convert a neutral utterance into an emotional utterance by incorporating features corresponding to each of the four components of speech either individually or in selected combinations. The resulting converted speech is compared with the actual emotional utterance to determine the closeness of the utterances, and hence the effect of changing the corresponding features.

It is also possible to remove the features selectively from the emotion utterance by transforming emotion utterance towards a neutral one. This way the significance of each of the four components in contributing to the emotion can be studied.

To modify the features of an utterance to the desired ones (neutral to emotion or emotion to neutral), a Flexible Analysis Synthesis Tool (FAST) is developed. The tool includes analysis of speech data to extract vocal tract system and excitation components using LP analysis [3]. The analysis also includes methods to extract the epochs (instants of significant excitation of vocal tract system) [6] and also the instantaneous fundamental frequency contour [7]. The key implementation tools are the DTW algorithm [2] for alignment of two utterances and a flexible prosody manipulation program.

Preliminary results of this study are presented in the form of effectiveness of the different components of speech that contribute to characterizing the emotional state of a speaker. The study is limited by the fact that the data is collected from the same speaker for both neutral and emotion conditions for the same sentence. Enacted data uttered by radio artists is used.

This paper is organized as follows. Section 2 describes the features related to the four components of speech used in this stduy. Methods used for extracting some of these features are also discussed briefly in this section. In Section 3 the design details of the flexible analysis synthesis tool (FAST) are described. The three major constituent blocks of the tool, namely, analysis, processing and synthesis are discussed. In Section 4 experiments conducted for emotion conversion (from neutral to emotion) are discussed along with observations and results. Finally, Section 5 gives a summary and scope for further studies.

2. Features for studying the characteristics of emotion in speech

Speech signal carries information about each of the four components – vocal tract system, excitation source, duration of sound units (syllables, words or phrases) and intonation. Choice of these four components of speech is based on the inherent characteristics of the production mechanism and also on the acquired characteristics such as prosody. They are not chosen for any of their emotion-specific features. In fact the objective of this study is to explore for any emotion-specific features in these components.

The first two components are related to the physical attributes of the speech production mechanism. The last two relate mainly to the behavioural traits of the speaker. Duration features are related to speaking style and speaking rate. Duration and intonation both are part of supra-segmental or prosody features. Comparison of utterances of the same sentence by a speaker in neutral mode and in emotion mode indicates variations in one or more of these components of speech [1]. Analysis of speech features corresponding to each of these components is useful to study the contribution of the components to emotion characteristics of speech.

In this paper, the following features of speech are examined:

- Vocal tract features to study the contribution of the system component
- Excitation features to study the contribution of the source component

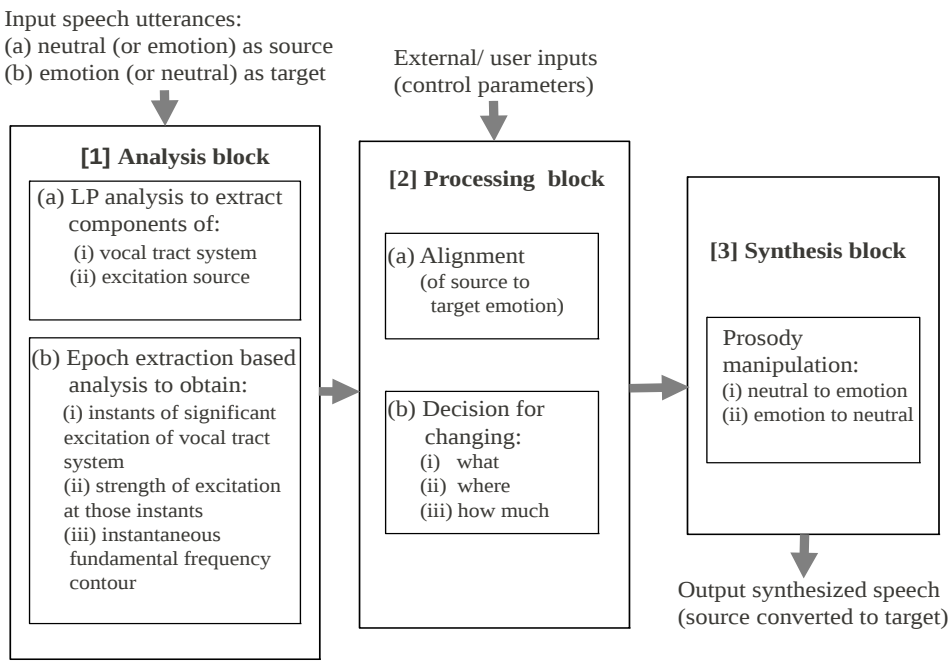

Figure 1: *Schematic block diagram of Flexible Analysis and Synthesis Tool (FAST) developed for studying the characteristic features of emotion in speech*

- Duration features to study the contribution of the supra-segmental component
- Instantaneous pitch contour features to study the intonation component of prosody

Synthesis of target (emotion) speech of the same sentence as that of the source (neutral) utterance is used to study the effect of different features. In the same way, synthesis by swapping these, i.e., synthesis of target (neutral) speech from source (emotion) utterance is also helpful in studying the significance of a feature on the emotion characteristics.

Vocal tract system features along with excitation components are extracted by linear prediction (LP) analysis. Linear prediction coefficients (LPCs) capture the information of the vocal tract shape in the form of spectral envelope. Linear prediction residual carries information of the excitation source. Parts of the LP residual of the source utterance are copied back to the corresponding regions around the new locations of epochs (instants of significant excitation), modified according to the target utterance [4].

Excitation source features, mainly the epoch locations i.e., instants of significant excitation of the vocal tract, are extracted using the zero frequency filtering (ZFF) method [6]. The method involves passing the differenced speech signal through a cascade of two zero frequency resonators, which have pair of poles on the unit circle in the z-plane at 0 Hz. The trend in the output is removed by subtracting the running mean computed over a frame size in the range of one to two (average) pitch periods. The resultant signal, i.e., ZFF signal, gives epoch locations at the instants of its negative to positive zero crossing locations. The slope of the ZFF signal around the epoch corresponds to the strength of excitation. These epoch locations obtained from source utterance are modified according to those of the target utterance. Modified instant locations are then utilized for synthesizing speech, by placing the original or modified LP residual around these instants.

Duration features of the entire source utterance are used, as no apriori knowledge is available about the relative significance of syllable, word or phrase level durations on the emotional content in speech.

The instantaneous pitch period is obtained by computing the intervals between successive epochs. Instantaneous pitch period contour of the source utterance is modified according to that of the target utterance by scaling each of its corresponding pitch periods. The prosody modified instantaneous pitch period contour is then used for synthesizing speech later.

3. Design details of Flexible Analysis and Synthesis Tool (FAST)

Two major objectives of the tool are (a) converting a neutral utterance into an emotional utterance by incorporating some features, and (b) converting an emotional utterance to neutral one by removing or modifying some features. The twin objectives can be achieved by facilitating conversion of a source utterance to a target utterance and synthesizing speech using the modified features. The three steps involved are: (i) analysis of speech to extract features corresponding to the four components of source speech, (ii) their alignment and modification according to features of the target speech, and (iii) synthesizing speech by conversion of the source utterance in accordance to the desired target utterance. The constituent blocks of the tool are shown in Figure 1.

The analysis block consists of LP analysis and epoch extraction. LP analysis is used for extracting the components of the vocal tract system and excitation source. Epoch extraction is used for obtaining the instants of significant excitation of the vocal tract (epochs), the strength of excitation at those instant

locations, and the instantaneous fundamental frequency contour (or instantaneous pitch period contour). In the analysis block extraction of these features and their subsequent analysis is carried out for source utterance and target utterance.

Time alignment of the source utterance with the target utterance is carried out using DTW algorithm [2]. The DTW algorithm consists of matching two speech patterns (X and Y), represented by the sequences of vectors $(\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_M})$ and $(\mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_N})$, where each vector $\mathbf{x_i}$ of pattern X and $\mathbf{y_j}$ of pattern Y correspond to the representations of the vocal tract shapes for the i^{th} and j^{th} frames, respectively. If LP analysis is used to capture the vocal tract shape information, then $\mathbf{x_i}$ and $\mathbf{y_j}$ may represent p-dimensional vectors such as weighted LP cepstral coefficients (wLPCC). Here M and N represent the durations of the speech patterns in terms of number of analysis frames in the patterns. The dissimilarity between the two patterns is obtained by using a distance measure between two vectors, and dynamic programming with constraints on the warping path [2].

The dynamic programming algorithm is based on the observation that the optimal path to the point $(i,\ j)$ in the two dimensional matrix must pass through either the point $(i-1,\ j)$, or $(i-1,\ j-1)$ or $(i,\ j-1)$. The minimum accumulated distance to point $(i,\ j)$ is then given by

$$D(i,j) = d(i,j) + min\{D(i-1,j), D(i-1,j-1), D(i,j-1)\} \quad (1)$$

where $d(i,\ j)$ is the distance between the pattern vectors $\mathbf{x_i}$ and $\mathbf{y_j}$. The algorithm recursively computes this distance to determine the minimum accumulated distance to the point $(M,\ N)$.

From the warping path it is possible to know the pitch period of the emotional (target) speech for each frame of the neutral (source) speech. Also, the warping path automatically provides the desired durational modification of the source utterance. Decision is also made regarding the features or parameters to be changed (what), their location on the time axis or duration (where), and the amount or degree of modification (how much) to be carried in those features or parameters. The external control inputs from user are used in this decision.

The synthesized speech is obtained through a flexible prosody manipulation program using the pitch and duration features [4]. Desired features of the source utterance are modified according to the features of the target utterance, the warping-path and the decision criterion. Modified features of the vocal tract (represented by LPCs) and excitation features (represented by LP residual) are added at new epoch locations. The tool facilitates the conversion both ways, i.e., neutral to emotion and emotion to neutral speech.

4. Experiments and Results

The emotion database named Simulated Emotion Speech Corpus collected by Indian Institute of Technology (IIT) Kharagpur (IIT-KGP SESC) is used for preliminary experiments conducted for designing and testing the tool. The database in Telugu (Indian) language consists of total around 12000 utterances spoken by radio artists. Each of the 10 speakers (5 male and 5 female) recorded utterances of 15 sentences, each spoken in 8 emotions and repeating these in 10 separate sessions of recordings. Hence utterances of a sentence from the same speaker are available in neutral mode as well as in 7 other emotions (anger, compassion, disgust, fear, sarcastic, happy and surprise).

In this study, 32 utterances of 4 speakers in neutral and in 7 emotions (anger, compassion, disgust, fear, sarcastic, happy and surprise) are taken from the database. Neutral (source) is coverted to emotion (target) speech. Prosody manipulation is carried out by modifying the duration (frame level) and intonation (pitch contour) features of the source (neutral) utterance.

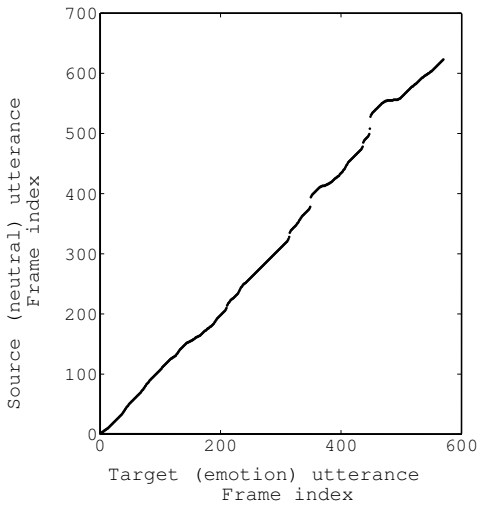

Figure 2: *Warping path (DTW) of target (emotion) utterance and source (neutral) utterance*

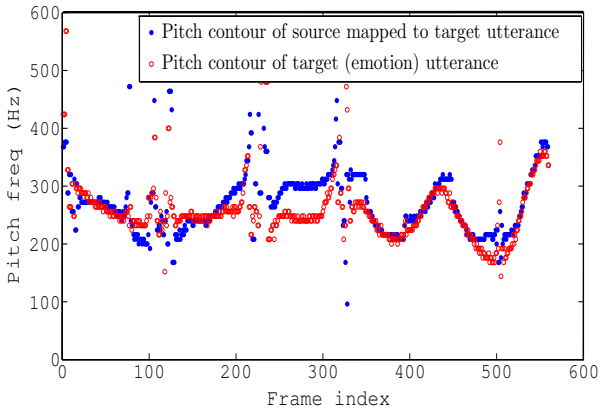

Figure 3: *Mapping of instantaneous pitch contour of source (neutral) utterance to target (emotion anger) utterance*

The analysis of source (neutral) and target (emotion) utterance is carried out in the analysis block, involving feature extraction by LP analysis and epoch extraction by ZFF.

The warping-path (i.e., time-alignment path) is obtained in the processing block using wLPCCs. An example warping-path obtained by the DTW alignment of source (neutral) utterance to target (emotion, for example anger) utterance is shown in Figure 2. Mapping of the instantaneous pitch contour of the source (neutral) utterance to target (emotion, for example anger) utterance, using the warping-path, is shown in Figure 3. The decision criterion for changing the prosody features (i.e., pitch and duration) is applied to decide the prosody manipulation parameters.

Synthesized speech is obtained in the synthesis block by conversion of the source (neutral) utterance in accordance with

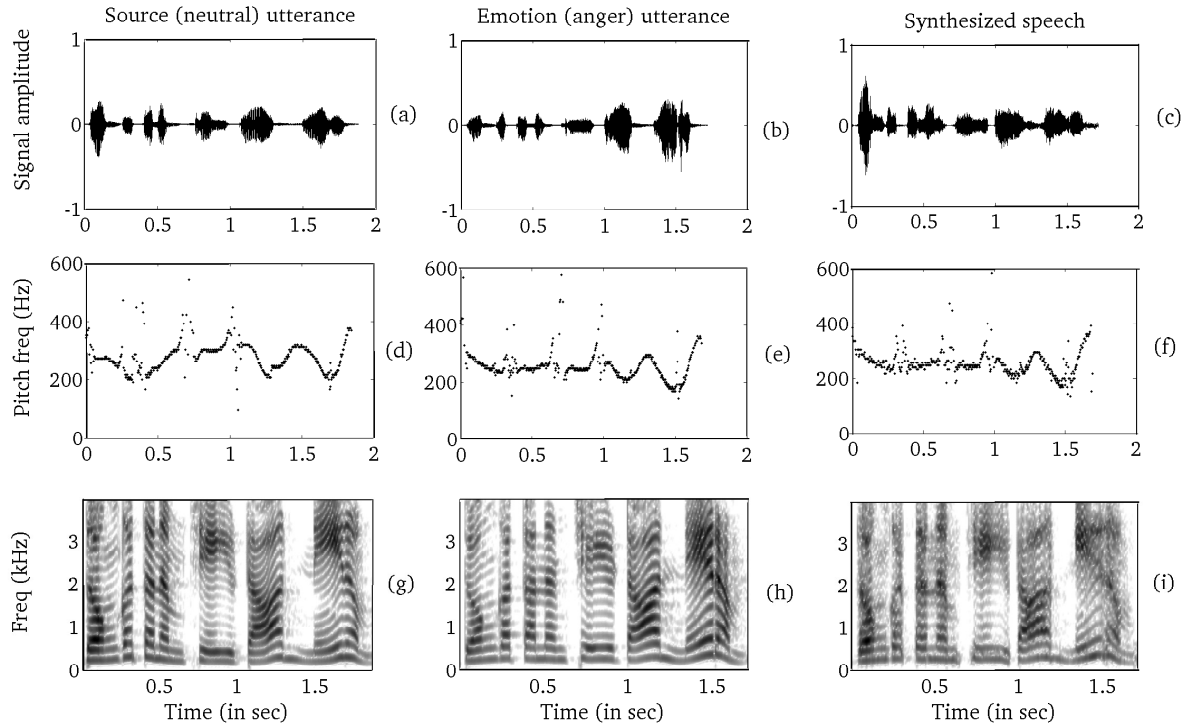

Figure 4: *Source (neutral) to target (emotion anger) conversion. Speech waveform, pitch contour and spectrogram of the source (neutral) utterance ((a), (d), (g)), target (emotion anger) utterance ((b), (e), (h)) and synthesized (emotion anger) speech ((c), (f), (i))*

Table 1: *Criterion for 'Similarity Score', used in perceptual listening test to judge the similarity/dissimilarity between two utterances.*

Perceptual differece between 2 utterances	Similarity score
sounds very much similar	5
sound some-what similar	4
sounds little different, little similar	3
sounds some-what different	2
sounds very much different	1

the features of the target (emotion) utterance. The warping-path and decision criterion obtained in the processing block are utilized in this conversion. The speech signal waveforms, pitch contours and spectrograms of source (neutral) utterance, target (anger) utterance and the synthesized emotion (anger) speech are shown in Figure 4.

It may be noted that the pitch contour of the synthesized speech (shown in Fig. 4(f)) bears a close resemblance to that of the target utterance (shown in Fig. 4 (e)). The spectrogram of the synthesized speech (shown in Fig. 4(i)) also shows good resemblance to that of the target utterance (shown in Fig. 4(h)). However some artifacts, possibly in the region of higher harmonics, can also be seen.

Subjective evaluation of the results is carried out through listening tests with 13 student listeners from the lab. Each subject was given 28 sets of utterances. Each set consisted of 3 files — one each of neutral (source) utterance, emotional (target) utterance and the synthesized speech. These 28 sets consist of utterances (drawn from the database [1]) of 4 different

Table 2: *Average similarity scores: for each emotion — between (a) source (neutral) and target (emotion) utterances, (b) source (neutral) utterance and synthesized (emotional) speech and (c) target (emotion) utterance and synthesized (emotion) speech.*

Emotion	(a)	(b)	(c)
1. Anger	2.48	2.36	3.77
2. Compassion	1.99	2.12	3.21
3. Disgust	2.34	2.21	3.38
4. Fear	2.07	1.98	3.50
5. Sarcastic	1.92	2.13	3.59
6. Happy	2.42	2.32	3.48
7. Surprise	2.07	2.15	3.63

speakers, each uttering a sentence in neutral mode and in 7 different emotions. The subjects were asked to judge the similarity/dissimilarity among the 3 files, and assign a similarity score for each of the 28 sets, on a scale of 1 to 5, as per the criterion shown in Table 1. Same criterion is used to assess the (i) similarity between source (neutral) and target (emotion) utterances, (ii) similarity between source (neutral) utterance and synthesized (emotion) speech and (iii) the similarity between target (emotion) utterance and the synthesized (emotion) speech.

Average similarity scores were then obtained for each speaker and for each emotion, for each of the following pairs: (a) source (neutral) and target (emotion) utterances, (b) source (neutral) utterance and synthesized (emotional) speech and (c) target (emotion) utterance and synthesized (emotion) speech. The average similarity scores for each of the 7 emotions are shown in Table 2 and Table 3, for each emotion

Table 3: *Average similarity scores*: *for each speaker − between (a) source (neutral) and target (emotion) utterances, (b) source (neutral) utterance and synthesized (emotional) speech and (c) target (emotion) utterance and synthesized (emotion) speech.*

Speaker : Emotion	(a)	(b)	(c)
Speaker 1: Anger	2.30	2.07	3.46
Speaker 1: Compassion	1.76	1.92	2.30
Speaker 1: Disgust	1.61	1.61	2.53
Speaker 1: Fear	1.53	1.69	2.92
Speaker 1: Sarcastic	1.30	1.69	3.53
Speaker 1: Happy	2.07	1.92	3.23
Speaker 1: Surprise	2.00	2.00	3.61
Speaker 2: Anger	2.23	2.15	3.38
Speaker 2: Compassion	1.61	1.69	3.07
Speaker 2: Disgust	2.23	2.23	3.46
Speaker 2: Fear	2.00	2.00	3.46
Speaker 2: Sarcastic	1.69	1.69	3.92
Speaker 2: Happy	2.30	2.07	3.38
Speaker 2: Surprise	1.84	2.00	3.61
Speaker 3: Anger	2.53	2.53	4.23
Speaker 3: Compassion	1.76	2.38	3.38
Speaker 3: Disgust	2.76	2.38	3.61
Speaker 3: Fear	1.92	2.07	3.38
Speaker 3: Sarcastic	2.07	2.69	3.23
Speaker 3: Happy	2.38	2.76	3.76
Speaker 3: Surprise	1.84	2.30	3.69
Speaker 4: Anger	2.84	2.69	4.00
Speaker 4: Compassion	2.84	2.46	4.07
Speaker 4: Disgust	2.76	2.61	3.92
Speaker 4: Fear	2.84	2.15	4.23
Speaker 4: Sarcastic	2.61	2.46	3.69
Speaker 4: Happy	2.92	2.53	3.53
Speaker 4: Surprise	2.61	2.30	3.61

and for each speaker, respectively.

The scores (refer column (c) in Table 2 and Table 3) indicate good similarity between the synthesized speech and target emotion utterance (for example − for anger, surprise, sarcastic and fear). It is especially true if the target utterance is carrying the emotion well, which is indicated by dissimilarity (low score of similarity) between the corresponding source (neutral) and target (emotion) utterances in column (a) in Table 2 and Table 3 (except for anger). Higher similarity scores (refer column (a)) between neutral (source) and emotion (target) utterances also indicate limitation of using the enacted data.

Some artifacts observed in the spectrogram of the synthesized speech (as shown in Fig. 4(i)) and lower speech quality of the synthesized emotion speech are due to that fact that only the duration and pitch contour features are modified in the current prosody manipulation. The features of the vocal tract system and excitation source are still that of the neutral (source) utterance. The target emotion is perceivable while listening to the synthesized speech.

5. Summary and scope of future work

The results of preliminary experiments show that it is possible to study the effectiveness of the different components of speech that contribute to a speaker′s emotional state. The key contribution of this work are the use of DTW algorithm for alignment of two utterances and a flexible prosody manipulation program developed to convert the emotional content in speech. It is envisaged that the flexibility of converting a neutral speech to emotional and vice-versa would be helpful to understand the components that contribute to emotion characteristics in speech. A vast variety of end user applications are possible, ranging from customized emotional consumer responses to customer expressive state analysis systems.

The current work is limited to efforts in conversion of neutral to emotion speech and limited subjective evaluation. Future work will include transformation of emotion utterance to neutral utterance by removing or modifying one or more features.

Acknowledgement

The authors would like to thank Prof. K. S. Rao of IIT, Kharagpur, India for sharing the valuable database IIT-KGP SESC.

This work is a part of ongoing research collaboration (2010-2014) on project Emotion between Speech and Vision Lab, IIIT, Hyderabad, India and SAIT, Samsung India Software Operations Pvt. Ltd., Bangalore, India.

References

[1] S. G. Koolagudi, S. Maity, V. A. Kumar, S. Chakravarti, and K. S. Rao, "IITKGP-SESC: Speech database for emotion analysis", *Communications in Computer and Information Science*, JIIT University, Noida, India: Springer, ISSN-1865-0929 ed., August 2009.

[2] H. Sakoe and S. Chila, "Dynamic programming algorithm and optimisation for spoken word Recognition", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 26, pp 43-49, 1978.

[3] K. S. Rao and B. Yegnanarayana, "Voice conversion by prosody and vocal tract modification", *International Conference on Information Technology*, print ISBN: 0-7695-2635-7, December 2006.

[4] K. S. Rao and B. Yegnanarayana, "Prosody modification using instants of significant excitation",*IEEE Transactions on Acoustics, Speech and Language Processing*, vol. 14, pp. 972-980, May 2006.

[5] D. Govind, S. R. M. Prasanna, and B. Yegnanarayana, "Neutral to target emotion conversion using source and suprasegmental information", *Proceedings of Interspeech 2011*, Florence, Italy, pp. 2969-2972, August 2011.

[6] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals", *IEEE Transactions on Acoustics, Speech and Language Processing*, vol. 16, no. 8, pp. 1602-1614, November 2008.

[7] B. Yegnanarayana and K. S. R. Murty, "Event based instantaneous fundamental frequency estimation from speech signals", *IEEE Transactions on Acoustics, Speech and Language Processing*, vol. 17, no. 4, pp. 614-625, May 2009.