Performance Analysis of Text To Speech Synthesis System using HMM and Prosody Features with Parsing for English Language

B.Sudhakar[1], R.Bensraj[2], C.R.Balamurugan[3]

[1]Assistant Professor, Dept. of ECE, Annamalai University

[2]Assistant Professor, Dept. of Electrical Engineering, Annamalai University

[3]Professor/EEE, Karpagam College of Engineering, Coimbatore

---***---

Abstract - This paper describes a Hidden Markov Model (HMM) based Text To Speech Synthesis (TTS) system and prosody based TTS system for producing natural sounding synthetic speech in English language. The HMM based system consists of two phases such as training and synthesis. English speech is first parameterized into spectral and excitation features using Glottal Inverse Filtering (GIF). An emotions present in the input text is modeled based on the parametric features. The performance measure has been carried out with recorded speech and the HMM based TTS system. Subsequently the TTS system with prosodic features for generating human voice has been implemented. To produce the output of TTS in the same form as if it is actually spoken the Prosody feature allows the synthesizer to vary the pitch of the voice. The pitch and duration play an important role to improve the naturalness of TTS output. The performance measure has been carried out with recorded speech and the Prosody based TTS system. Finally the performance of HMM based TTS has been compared with Prosody based TTS to measure the effectiveness of the system. Both TTS systems are used to analyze the emotions such as Happy, Fear, Neutral and Sad to improve the effectiveness of the system.

Key Words: Hidden Markov Model (HMM), Glottal Inverse Filtering (GIF), Text- To- Speech (TTS).

1. INTRODUCTION

The HMM is an effective technique for modeling the acoustics of speech and it has enabled significant progress in speech and language technologies [1,2]. It is a statistical model used more often for speech synthesis. A basic block diagram of HMM based speech synthesis consists of training and synthesis phase. In the training phase speech signal is parameterized into excitation and spectral features. The HMM is trained using these features. In the synthesis phase, given text is transformed into a sequence of context dependent phoneme labels. Based on the label sequence, a sentence HMM is constructed by concatenating context-dependent HMMs. From the sentence HMM, spectral and excitation parameter sequences are obtained. It is synthesized through interpolating and concatenating natural pulses, and the excitation signal is further modified according to the spectrum of the desired voice source characteristics. Speech is synthesized by filtering the reconstructed source signal with the vocal tract filter. HMM based speech synthesis has many attractive features such as complete data driven voice building, flexible voice quality control, and speaker adaptation.

The major advantage of HMM based speech synthesizers is their higher parametric flexibility [3,4]. It is also used to transform voice characteristics, e.g. specific voice qualities and basic emotions. The main characteristics of these systems are High-quality speech and robustness to variations in speech quality. Fully parametric. Fully automatic. Easy to transform voice characteristics. New languages can be built with little modification. Speaking styles and emotions can be synthesized using a small amount of data. These characteristics make this technique very attractive, especially for applications which expect variability in the type of voice and a small memory footprint [5, 6].

Prosody refers to the characteristics of speech that make sentences flow in a perceptually natural, intelligible manner. Without these features, speech would sound like a reading of a list of words. The major components of prosody that can be recognized perceptually are fluctuations in the pitch, loudness of the speaker, length of syllables, and strength of the voice. These perceptual qualities are a result of variations in the acoustical parameters of fundamental frequency (F0), intensity (amplitude), phonemic duration, and amplitude dynamics.

The parsed text with a phoneme string is the input of prosody generator. The input text is broken into prosodic phrases, possibly separated by pauses, and assigning labels to different syllables or words within each prosodic phrase. The words are normally spoken continuously, unless there are specific linguistic reasons to signal a discontinuity. The term juncture refers to prosodic phrasing that is, where do words cohere, and where do prosodic breaks (pauses and/or special pitch movements) occur.

The primary phonetic means of signaling juncture are:

i. Silence insertion.

ii. Characteristic pitch movements in the phrase-final syllable.

iii. Lengthening of a few phones in the phrase-final syllable.

iv. Irregular voice quality such as vocal fry

The duration of each phoneme, volume and the pitch contour is delivered by the prosody generator. Prosodic features (i.e. intonation and phonemic duration) are determined based on the phrase accents, syllabic accents, and phoneme location. These features are represented by actual pitch and duration values for each phoneme. Accurate determination of the pitch and duration values is essential for producing more natural sounding speech.

2. HMM BASED TTS SYSTEM

The proposed HMM based English TTS system targets to produce natural sounding synthetic speech capable of carrying diverse emotions. To accomplish this objective, the task of the real human voice construction machine is modeled by utilizing GIF or Glottal source modeling entrenched in an HMM framework [7 ,8, 9].

The motivations to use glottal source modeling in HMM-based speech synthesis are:

Reduce business of synthetic speech.

Better modeling of prosodic aspects which are related to the glottal source.

Control over glottal source parameters to improve voice transformations

2.1. Parameterization Phase

To eliminate the possible low frequency fluctuation from the signal, the signal is high pass filtered in the parameterization stage. The signal is then windowed with a rectangular window in 25-ms frames at 5ms. The log energy is estimated from the windowed speech signal. Then GIF is accomplished in order to evaluate the glottal volume velocity waveform from the speech signal. Iterative

Adaptive Inverse Filtering (IAIF) is employed for the automatic GIF. It repeatedly withdraws the effects of the vocal tract and the lip radiation from the speech signal using all-pole modeling. The assessed glottal flow signal and the Linear Predictive Coding (LPC) model of the vocal tract are the outputs of the inverse filtering block. In order to arrest the deviations in the glottal flow due to different phonation or speaking style, the spectral envelope of the glottal flow is further parameterized with LPC.

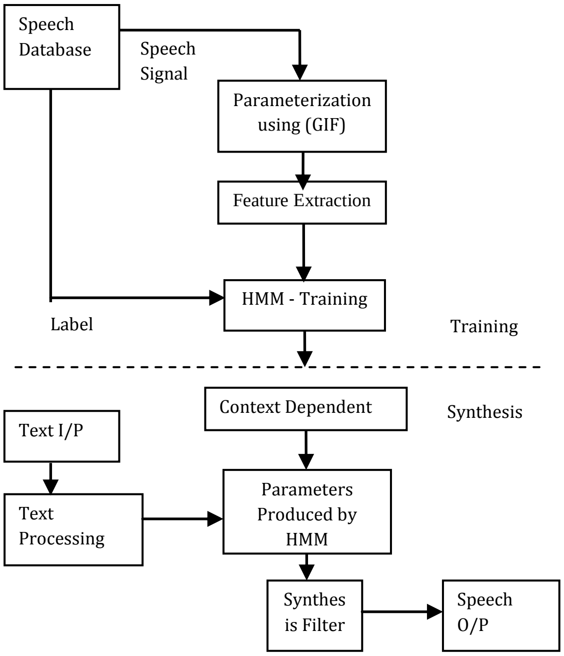


Figure : 1 A Basic Block Diagram of HMM Based Speech Synthesis System

2.2 HMM based TTS system using GIF

The speech is first decomposed into the glottal source signal and the model of the vocal tract filter through glottal inverse filtering. The most common approach for models of speech production separates the process into three distinct steps: Glottal Excitation G(z), Vocal Tract Filtering V(z) and Lip Radiation L(z). From the speech pressure signal the spectral and glottal excitation is computed using GIF. The below equation is used to estimate the glottal excitation G(z).

$$G(z) = \frac{S_S(z)}{V(z)\,Lr(z)}$$

where

Ss(z)	=	Z-Transform of Speech signal
Vt(z)	=	Z-Transform of Vocal tract
Lr(z)	=	Z Transform of Lip radiation effect

Parametric feature expression in the voice source and the vocal tract transfer function are computed in the parameterization phase using automatic GIF. The Vocal tract transfer function VTTF normally consists of poles and zeros and can we expressed as

$$V(z) = \frac{b_0 \prod_{k=1}^{m}(1 - d_k Z^{-}1)}{\prod_{k=1}^{n}(1 - C_k Z^{-}1)}$$

where b_0, c_k and d_k represent gain factor, poles of V(z) and zeros of V(z). The poles represent several peaks associated with the resonance of the acoustic cavities that from the vocal tract. These resonances are measured by formants. Each formant is described by its formant frequency and its formant bandwidth. The zeros or antiresonance of the VTTF represents energy loss and located at very high frequencies.

The proposed system comprises of two main parts: training and synthesis shown in Figure 1. In the training phase, spectral parameters, namely, Mel Cepstral coefficients and their dynamic features, the excitation parameters, namely, the log fundamental frequency (F0) and its dynamic features, are extracted from the speech data using GIF. The HMM is trained using these features. In the synthesis phase, first, an arbitrarily given text is transformed into a sequence of context dependent phoneme labels. Based on the label sequence, a sentence

HMM is constructed by concatenating context-dependent HMM. From the sentence HMM, spectral and excitation parameter sequences are obtained based on the Mel Log criterion. The context-dependent phone models are used to capture the phonetic and prosody co-articulation phenomena. Finally, vocoder speech is synthesized from the generated spectral and excitation parameter sequences by concatenating context-dependent HMM. It is used to analyze the emotions such as Happy, Fear, Neutral and Sad to measure the effectiveness of the system.

3. TTS SYSTEM USING PROSODY FEATURES WITH PARSING

This TTS system is composed of two parts a front end and a back end. The front end has two major tasks. First, it converts raw text containing symbols like numbers and abbreviations into the equivalent of written out words. This process is often called text normalization, preprocessing, or tokenization. The front end, then assigns phonetic transcriptions to each word, and divides and marks the text into prosodic units, like phrases, clauses, and sentences. The process of assigning phonetic transcriptions to words is called text-to-phoneme conversion. Phonetic trans criptions and prosody information together make up the symbolic linguistic representation that is output by the front end. The back end often referred to as the synthesizer then converts the symbolic linguistic representation into sound (10, 11,12).

Figure 2. presents a general block diagram of the TTS synthesis system using prosody features like pitch, pause, stress, phoneme duration, etc.,. First the incoming text must be accurately converted to its phonemic and stress level representations. This includes determination of word boundaries, syllabic boundaries, syllabic accents, and phoneme boundaries. The text preprocessing finds the word boundaries. Subsequently the prosodic parsing involves the determination of phrase boundaries and phrase level accents. Parsing is a method of scanning the text, in order to determine various points such as content of text, context of text, frequency of particular word in the text, etc. While finding out the emotions present in the text, it is necessary to determine the context of the text. The context of the text determines the current emotions present in the text and also used to find variation in the emotion. A database is maintained, which contains the keywords and category of emotion to which it belongs. The text is scanned and keywords present in the text are compared with the contents of the database. The comparison will finalize the value of emotion. The various emotions such as happy, fear, neutral and sad have been analyzed by this system. The prosody generator finds the prosodic features such as pitch, duration and intonation, etc, related to the emotions. Then the system produced appropriate emotions for the text input. Finally concatenation is carried out to produce the synthesized speech output.

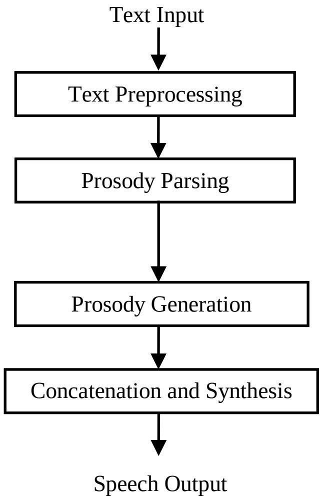
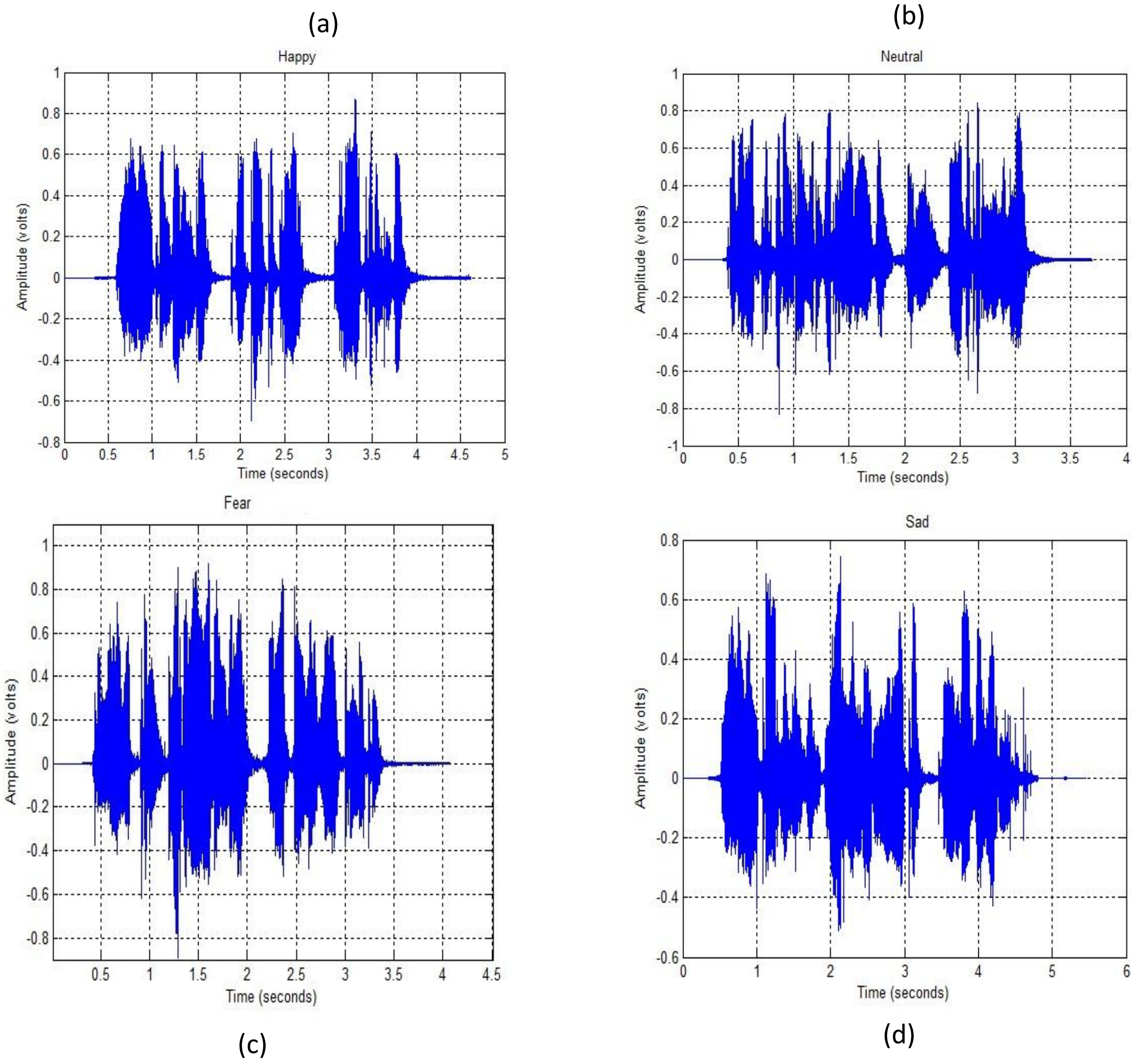
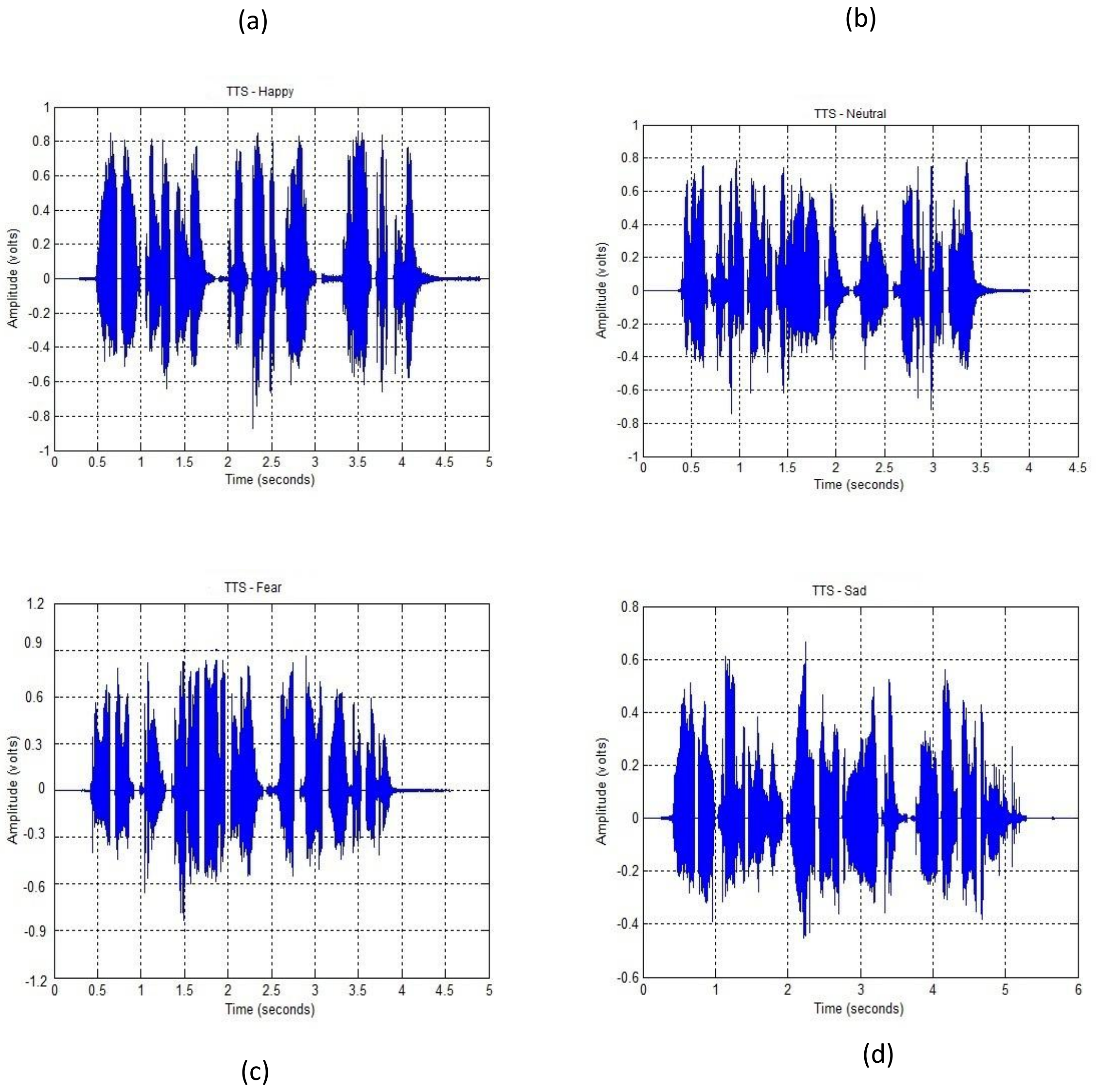


Figure 2. General Block Diagram of TTS Synthesis System using Prosody features

4. RESULTS AND DISCUSSION



Figures 3 (a-d). Recorded voice of different emotions



Figures 4 (a-d). HMM based TTS output of different emotions

The figures-3 (a,b,c,d) show the wave form of different emotions of recorded speech in the noise free environment Table-1 illustrate the different emotional sentences used by the TTS systems and recorded speech. The recorded speech features are taken from Table-2 which are used as the reference for analyze the performance of this system. Experimental results of HMM based TTS for different emotions are illustrated in Figures 4 (a,b,c,d). The performance analysis has been carried out through compare the recorded speech with HMM based TTS system to measure the naturalness achieved in this system.

Table 1. The different emotional sentences used by the TTS system and recorded speech

Emotions	Recorded Sentences
Happy	I got 99 out of 100 in my maths subject.
Fear	I am realy afrai about my Tamil subject.
Neutral	Tomorrow I will go to my school.
Sad	I got 10 out of 100 in my chemistry subject.

Table 2: Amplitude variation and spectral mismatch of recorded speech

Emotions	Amplitude (V)	Spectral Mismatch (S)
Happy	0.95	0
Fear	0.91	0
Neutral	0.83	0
Sad	0.70	0

Table 3. Amplitude variation and spectral mismatch of Prosody based TTS output

Emotions	Amplitude (V)	Spectral Mismatch (S)
Happy	0.875	0.25
Fear	0.830	0.270
Neutral	0.746	0.284
Sad	0.645	0.295

Table 4. Percentage of naturalness achieved in Prosody based TTS output

Emotions	Amplitude (V) (%)	Spectral Mismatch (S) (%)	Average Naturalness achieved (%)	Mean Average of all emotions (%)
Happy	93.7	74.0	83.2	82.38
Fear	91.2	73.0	82.54	
Neutral	90.21	71.34	80.85	
Sad	82.8	70.5	76.65	

Table 5. Amplitude variation and spectral mismatch of HMM based TTS output

Emotions	Amplitude (V)	Spectral Mismatch (S)
Happy	0.91	0.013
Fear	0.882	0.043
Neutral	0.772	0.061
Sad	0.671	0.062

Table 6. Percentage of naturalness achieved in HMM based TTS output

Emotions	Amplitude (V) (%)	Spectral Mismatch (S) (%)	Average Naturalness achieved (%)	Mean Average of all emotions (%)
Happy	97.3	95.9	95.6	94.8
Fear	96.6	95.4	94.9	
Neutral	95.4	94.8	93.9	
Sad	89.4	93.8	91.6	

Table- 3 shows the amplitude variations and spectral mismatch of the prosody based TTS system. A comparative performance measure has been carried out with Table- 2 and Table- 3 for finding the variations of emotional features. The percentage of naturalness achieved for individual emotions are listed in Table- 4 . It shows that the sentence related to happy emotion is achieved highest rating and the lowest rating is related to sad emotional sentence. The mean average of naturalness achieved by all emotions is used to evaluate the overall performance of the system. An evaluation of mean average naturalness achieved for all emotions shown in Table-4 projected that the prosody based system is achieved 82.38% naturalness.

Table- 5 shows the amplitude variations and spectral mismatch of the HMM based TTS system. A comparative performance measure has been carried out with Table- 2 and Table- 5 for finding the variations of emotional features. The percentage of naturalness achieved for individual emotions are listed in Table- 6. It shows that the sentence related to happy emotion is achieved highest rating and the lowest rating is related to sad emotional sentence. The mean average of naturalness achieved by all emotions is used to evaluate the overall performance of the system. An evaluation of mean average naturalness achieved for all emotions shown in Table-6 projected that the HMM based system is achieved 94.8% naturalness.

5. CONCLUSION

The HMM based TTS system and prosody based TTS system has been developed for the English language. The emotional speech has been generated by HMM using the parametric features of GIF. The prosody feature allows the synthesizer to vary the pitch of voice to produce the output of TTS in the same form as if it is actually spoken. The emotions such as happy, fear, neutral and sad are analyzed to measure the effectiveness of the HMM based TTS and prosody based TTS systems. The performance analysis has been carried out for both systems to measure the naturalness achieved by the individual system. Based on this analysis the HMM based system produced highest naturalness (94.8%) than a prosody parsing based system (82.38%). Experimental results show that the HMM based TTS system is accomplished by generating natural sounding speech, and the quality is obviously better.

6. REFERENCES

[1] R. D.Novan and P. Woodland, "A hidden markov-model-based trainable speech synthesizer", *Computer Speech Language*, Vol. 13, No. 3, pp. 223–241,1999.

[2] D.Jurafsky and J.H.Martin, "Speech and language processing", *Pearson Education*, India, 2000.

[3] T. Oshimura and T. S. Kitamura, "Simultaneous modeling of spectrum pitch and duration in HMM-based speech synthesis", in *IEICE Transactions*, pp. 2099–2107, 2000.

[4] G.L. Jayavardhana Rama, A.G. Ramakrishnan, "A complete text-to-speech synthesis system in tamil", in *Proceedings of IEEE Workshop on Speech Synthesis*, pp. 191-194, 2002.

[5] M.B.Changak and R.V.Dharskar, "Emotion extractor based methodology to implement prosody features in speech synthesis", *International Journal of Computer Science*, Vol. 08 , No.11 , pp. 371–376, 2011.

[6] V.Francisco, R.Hervas, F.Peinado, and P.Gervs, "EmoTales: Creating a corpus of folk tales with emotional annotations", *Language Resource and Evaluation*, Vol. 46 , No.03 , pp. 341–381, 2012.

[7] Vibavi Rajendran and G.Bharath, "Text Processing for Developing Unrestricted Tamil Text to Speech Synthesis System", *Indian Journal of Science and Technology*, Vol. 08, No.29, pp .112-124, 2015.

[8] Sangaransing Kayty and Monica Munda, "A marathi HMM based speech synthesys system", *Journal of VLSI and Signal processing* , Vol. 05, No.06, pp .34-39, 2015.

[9] VA.Natarajan and S.Jothilakshmi , " Segmentation of continuous Tamil speech into syllable like units", *Indian Journal of Science and Technology,* Vol. 08, No.17, pp.417–429, 2015.

[10] Burkhardt, F., & Sendlmeier, W. F., Verification of Acoustical Correlates of Emotional Speech using Formant-Synthesis, ISCA Workshop on Speech &Emotion, Northern Ireland, pp. 151-156, 2000.

[11] Vroomen,J.,Collier,R.,&Mozziconacci,S.J.L., Duration and Intonation in Emotional Speech, Europeech, Vol. 1, pp. 577-580, 1993.

[12] He,L.Huang,,H.and Lech,M,. Emotional speech synthesis based on prosodic feature modification., Engineering, Vol.5, pp.73-77, 2013.