ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

Enhanced Sentimental Analysis for Speech Synthesis Based on Prosody Feature Modification Using TD-PSOLA Technique for Tamil

B. Sudhakar[1], R. Bensraj[2], J.Sangeetha[3]

Assistant Professor, Department of Electrical Engineering, Annamalai University, Chidambaram, India[1]

Assistant Professor, Department of Electrical Engineering, Annamalai University, Chidambaram, India[2]

Ph.D Scholar, Annamalai University, Chidambaram, India[3]

ABSTRACT: In recent years the synthesis of sentimental speech has various applications in customer services in the area of analyzing business adds in social media, mobile services and human- computer interaction etc. An enhanced sentimental analysis for Tamil text to speech (TTS) synthesis systems are produced by modifying prosody feature using Time Domain Pitch Synchronous Over Lap Add (TD-PSOLA) technique has been proposed in this paper. The existing speech synthesis systems have lack of naturalness in output speech and emotions. For experiment analysis four various types of emotions has been produced. The experiment results shows that the naturalness of synthesized output has been enhanced in proposed emotional speech synthesis system to achieves an enhanced performance.

KEYWORDS: Time Domain Pitch Synchronous overlap Add (TD-PSOLA) ,Tamil Text to Speech (TTS), Sentimental Speech Synthesis (SSS)

I. Introduction

Over the past years TTS system has become one of the most important research area due to its important in various applications. Text to speech synthesis is the process of converting normal text to speech signal [1]. In the call centers the speech synthesizer could conduct conversation with customers. The intelligent virtual agent devices or machine could read loud to users using speech synthesis techniques, such as in video games or in children toys. An important disadvantage existing in the machines could not express with human sentiments Sentiment analysis is the natural language processing task dealing with sentiment detection and classification from texts. Sentiment analysis has become one of the essential field fo29392939 research in the computational linguistics and is used or analyzing the people’s expression in either speech or text. The ultimate motive of sentimental speech synthesis (SSS) has to produce the natural speech as human voice. There are two main methods are used for speech production. These methods are formant synthesis and concatenation synthesis is illustrated in [2]. The formant synthesizer utilizes a simple model of speech generation and a set of rules to generate the speech. But the formant synthesis is not frequently utilized, because the resulting speech has lack of perfection. Alternatively concatenative synthesis links recordings of human speaker to generate the synthetic speech [3]. The generating utterance are more natural. In order to produce diversified emotions (or) sentiments the system needs a large size of data base. To solve this problem researchers found prosodic features extraction to enhance the naturalness of sentimental output [4,5].In this aspect very few numbers of speech corpora is needed. Diversified types of sentiments could be liked in to synthesized speech by changing appropriate to acoustic parameters either fundamental frequency or the speech contour duration, and then incorporate the concatenative approaches such as PSOLA (Pitch Synchronization Overlap Add) technique [6]. This paper proposed an enhanced sentimental analysis for TTS synthesis systems are produced by modifying prosody feature using TD-PSOLA technique has been implemented.

DOI: 10.15680/IJIRCCE.2016. 0403008

II. PROSODY FEATURES

There are various prosodic features available in the speech signals ,such as pause, pitch, stress, volume, time duration, fundamental frequency Fo, and energy.

A. Pause:

It is a temporary stop (or) inaction especially as caused by uncertainty (or) hesitation is a non-fluency feature. However, intentional pauses are used to demarcate units of grammatical construction, such as sentences or clauses. These can be indicated in writing by full stops, colons, semi-colons and commas.

B. Pitch:

It is the highness or lowness of a tone as perceived by the ear, which depends on the number of vibrations per second produced by the vocal cords. Pitch is the main acoustic correlate of tone and intonation. Different pitch levels, or intonation, can affect meaning.

C. Stress:

It is the relative emphasis that may be given to certain syllables in a word, or to certain words in a phrase or sentence. It is typically signaled by such properties as increased loudness and vowel length, full articulation of the vowel, and changes in pitch .Stress, or emphasis, is easy to use and recognize in spoken language, but harder to describe. A stressed word or syllable is usually preceded by a very slight pause, and is spoken at slightly increased volume

D. Volume:

Apart from the slight increase in loudness to indicate stress, volume is generally used to show emotions such as fear or anger. In writing, it can be shown by the use of an exclamation mark, or typographically with capitals or italics (or both)

E. Time duration:

The speech signals are having onset and an offset, and physical duration is determined by the time interval elapsing between onset and offset. This duration is duration is calculated by perceived onset and perceived offset, and an appropriate perceptual measure of time elapsed. This is not only so for the duration of sounds, but also for the duration of silent intervals between sounds.

F. Fundamental frequency Fo:

It is a lowest frequency which is produced by the speech signal, as distinct from the harmonics of higher frequency.

G. Energy:

It is an another important prosodic feature to produce the sentimental TTS system output from the speech signal.

III. SENTIMENTAL TTS SYSTEM

The prosody features are modified based on sentimental TTS has been proposed in this paper. The sequential process of sentimental TTS system based on prosody feature modification using TD-PSOLA concatinative technique is illustrated in fig.1.In this paper ,the speech utterances, which are retrieved from sentimental speech database with four diversified types of emotions anger, joy and sadness are split in to a set of units. For the concatinative speech synthesis there are multiple choices for the type of unit .The prominently utilised types includes words, syllables, phonemes and diphones. The basic synthesis building block is the speech unit. When the unit length is longer the synthesis speech quality will increase. In this paper, considered the size of speech database so the selected concatinative unit is word .The pitch, energy and duration rules are enumerated using prosodic analysis for each segments. To decide the pause assignment the length of silence is also calculated. Using the estimated parameters the prosody feature templates for diversified emotions are generated.

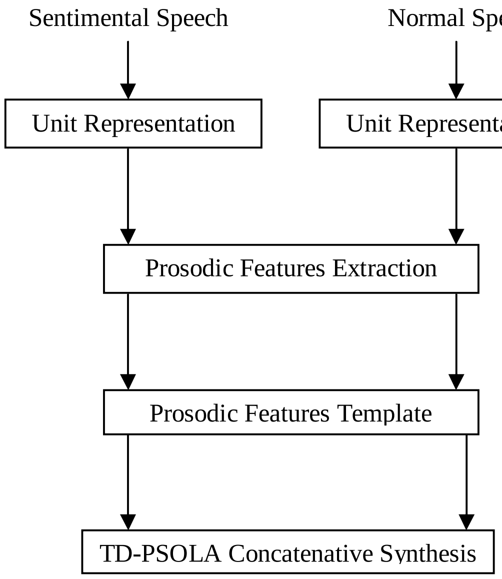
Sentimental Speech
Normal Speech
Unit Representation
Unit Representation
Prosodic Features Extraction
Prosodic Features Template
TD-PSOLA Concatenative Synthesis

Figure .1. Sentimental speech synthesis flow diagram

For the normal input speech, initially the utterance are segmented in to word or units. Then the prosody features are extracted for each unit, and modified according to prosody feature templates with the related sentiment types. Finally the synthesised sentimental output will be generated through TD-PSOLA concatinative technique, which is used to smooth and modify the boundary unit.

IV. GENERATION OF TTS USING PROSODIC FEATURES

The TTS system generated by different elements of prosodic features are illustrated in Figure2.The Tamil text message with a phoneme string is the input of prosody generator. The duration of each phoneme and the pitch contour is delivered by the prosody generator. Before applied input to the Prosody generator the input is converted in to phonemes based on the key stokes involved in the characters present in the input. The duration and pitch of each phoneme depends on the content and context of the text. In the context, the sentiment of conversation is sad, then pitch of word is altered accordingly to allow listener to understand sad sentiments of content. So prosody has vital roll in guiding listener recovery of the basic messages.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

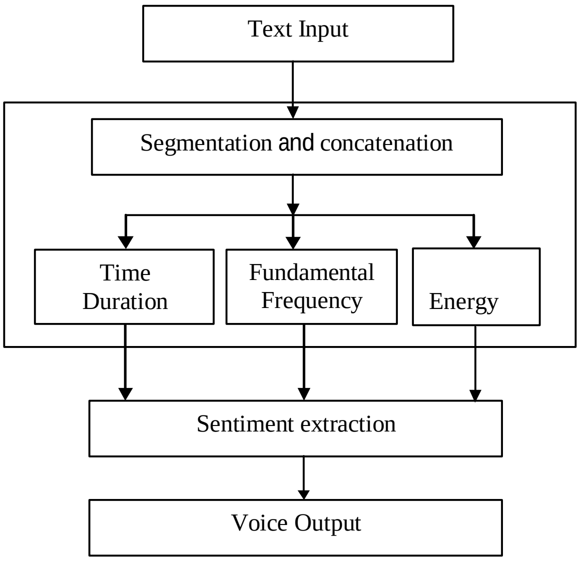

Figure.2. TTS system Using Prosodic Features

V. VARIOUS PHASES OF PROSODY GENERATOR

5.2. Speaking Style

Prosody depends not only on the linguistic content of a sentence. Different people generate different prosody for the same sentence. Even the same person generates a different prosody depending on his or her mood. The speaking style of the voice in Figure 2, can impart an overall tone to a communication. Examples of such global settings include a low register, voice quality (falsetto, creaky, breathy, etc.,), narrowed pitch range indicating boredom, depression, or controlled anger, as well as more local effects, such as notable excursion of pitch, higher or lower than surrounding syllables, for a syllable in a word chosen for special emphasis. The various parameter which influence the speaking Style are presented in [8,9].

5.2. Character

It is an important determining element in prosody, refers primarily to long-term, stable, extra-linguistic properties of a speaker, such as membership in a group and individual personality. In determines the features such as gender, age, speech defects, etc. affect speech, and physical status may also be a background determiner of prosodic character. Finally, character may sometimes include temporary conditions such as fatigue, inebriation, talking with mouth full, etc. Since many of these elements have implications for both the prosodic and voice quality of include temporary conditions such speech output, they can be very challenging to model jointly in a TTS system.[10,11]

5.3. Emotion*:*

The prosody should consider the temporary emotional conditions such as amusement, anger, contempt, sympathy, suspicion, etc., are suggested in[8,9]. A large number of high-level factors go into determining emotional effects in speech. Among these are point of view (can the listener interpret what the speaker is really spontaneous vs. symbolic (e.g., acted emotion vs. real feeling); culture-specific vs. universal; basic emotions and compositional emotions that combine basic feelings and effects; and strength or intensity of emotion.

Anger*:* It may be too broad a category for coherent analysis. One could imagine a threatening kind of anger with a tightly controlled F0, low in the range and near monotone; while a more overtly expressive type of tantrum could be correlated with a wide, raised pitch range.

Joy: It is generally related with increase in pitch and pitch range, with increase in speech rate. Untrained listener can easy to identify smiling because it raises F0 and formant frequencies.
Sad: It has normal or lower than normal pitch realized in a narrow range, with a slow rate and tempo. It may also be characterized by slurred pronunciation and irregular rhythm.
Fear: It is characterized by high pitch in a wide range, variable rate, precise pronunciation, and irregular voicing.
In this work the conventional Berlin Emotional Speech (BES) database is used for the purpose of comparison with other experiments [7].The test samples are carefully taken to produce naturalness while construct the database. Totally 8 frequently used sentences (4 long sentences and 4 short sentences) selected from every day communication. The speech utterance are produced by same speaker. For each of four emotions (angry, joy, sadness and fear), an entire unit selection database was recorded by the same speaker. Totally 300 sentimental outputs have been recorded to construct the sentimental prosodic feature templates.

VI. PROSODIC FEATURES CALCULATION

Among the different prosodic features three features are taken for calculations, which are time period, fundamental frequency and energy.

6.1 Calculation of fundamental frequency

The fundamental frequency F0 speech signal can be calculated either by time domain using the autocorrelation method [12],(or) frequency domain using the cepstral method [12]. The ultimate value of F0 can be calculated by average these two measurements. The fundamental frequency F0 can directly calculated through time domain autocorrelation method. The waveform utilizing the autocorrelation function which is expect to show peaks at delays corresponding to multiples of the glottal wave period (1/F0). The autocorrelation is calculated as

$$R_n(j) = \sum^{N-1-j} s_n(i)\, s_n(i+j)$$

where s is the speech signal. ⟶ (1)

The frequency domain cepstrum approach is used to calculate the F0 indirectly. It expects for a periodicity in the log spectrum of the speech waves ; if the log magnitude spectrum contains many regularly spaced harmonics, then the Fourier analysis of the spectrum will expose a peak corresponding to the spacing between the harmonics: i.e. the fundamental frequency.

6.2 Calculation of Energy

Speech signal is a non-stationary time varying signal. However, it could be viewed as a stationary signal in a short span ranging between 15 ms and 30ms.In the experiment, the short span energy is calculated

$$E_{\hat{g}} = \sum_{p=\hat{g}-J+1}^{\hat{g}} (S[p]h[\hat{g}-p])^{2} \longrightarrow (2)$$

where s[p] is the speech signal, h[$\hat{g}$ −p] is the applied window. $\hat{g}$ = tT , where T represents frame shift and t is the integer.

6.3 Calculation of time duration.

To generate the prosodic characteristics of speech signals the time duration of each unit under diversified sentimental states are calculated, in order to produce the pause assignment for an individual sentiments the duration of silence in each sentence has been estimated. The speech endpoint detection technique is used to detect the duration of silence in the speech. The ultimate aim of endpoint estimation algorithm is to detect the speech signal from background noise. The short-time energy is enumerated to explore voiced speech and short-time zero crossing rate is calculated to decide the voiceless speech [13]. The speech parts will be eliminated to calculate the length of silence.

Table 1 - Mean values of Prosodic features for different sentiments

Prosodic Features	Frequency Fo (HZ) x 10^2	Period (S)	Energy (dB) x 10^2
Angry	2.75	0.21	0.821
Joy	2.21	0.22	0.713
Sad	1.51	0.35	0.601
Fear	1.71	0.26	0.572

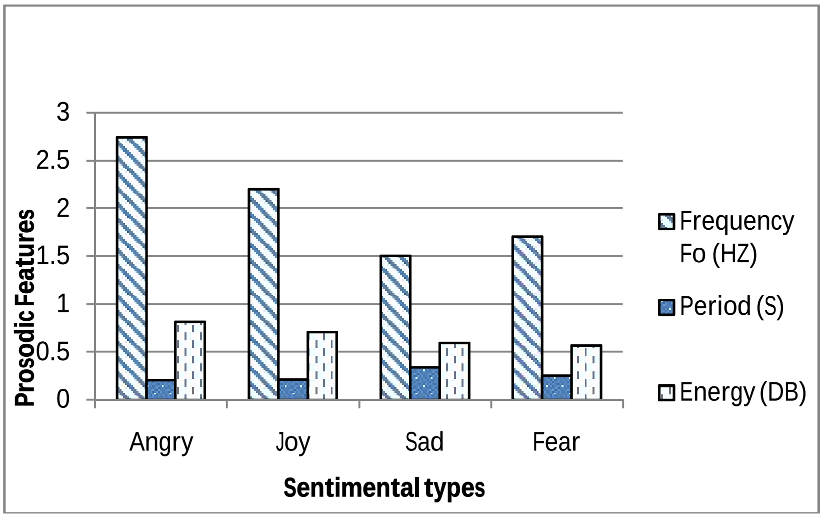

Figure. 3 . Performance analysis of Prosodic Features for different sentiments

VII. TD-PSOLA TECHNIQUE

TD-PSOLA is a prominently used concatenative synthesis technique to produce naturalness in the generated speech signal.The ultimate aim of TD-PSOLA technique is to alter the pitch directly on the speech waveform. The TD-PSOLA technique sequentially follows three steps, which are pitch synchronization analysis, pitch synchronization modification and pitch synchronization synthesis. Pitch synchronization analysis plays an important roll of TD- PSOLA technique, it executes two tasks: fundamental frequency detection and pitch mark. Let $X_n(m)$ denotes the windowed short time signal:

$$x_n(m) = h_n(t_n\text{-m})x(m) \longrightarrow \quad (3)$$

where $t_n$is the mark point of pitch, h_n is the window sequence.

Pitch synchronization modification links the pitch mark by changing the duration (insert or delete the sequence with the length of pitch duration) and tone (in- crease or decrease the fundamental frequency). The pitch synchronization synthesis adds the new sequence signal produced in the previous step

$$\bar{X}(\mathrm{m}) = \frac{\sum_{i} b_j\,\bar{x}_j(\mathrm{m})\bar{h}_j(\bar{t}_j - \mathrm{m})}{\sum_{j} \overline{h^2}_j\,(\bar{t}_j - \mathrm{m})} \qquad (4)$$

Where $\bar{t}_j$ is the new pitch mark, $\bar{h}_j$is the synthesized window sequence, $b_j$is the weight to compensate the energy loss when modifying the pitch value.

VIII. RESULTS AND DISCUSSION

The results generated from the proposed system has enhance the naturalness in the speech signal. Figure 4 shows the waveforms of the utterance “India Srilankavidam Vetri Petrathu” produced under four types of emotional states: angry, joy, sad and fear**.** Figure.5 also shows the waveforms of the synthesized sentimental speech under four different types of emotional conditions based on the prosodic feature modification algorithm. The Figure.4 disclose the simulated waveforms of the speech signals pronounced by human beings under four diversified types of sentiments. It is obviously shows that the waveforms of the synthesized speech from Figure.5 are eminent among different types of sentiments and they are similar to the waveforms of natural speech which is produced at Figure.4. From Table 2**,** it is easy to segregate sentiment types from the synthesized utterances by human being. The sentiment “angry” is easiest to identify from the synthesized speech. This is because the natural emotion “angry” contains physically powerful emotional provocations, resulting in notable prosodic characteristics. The sentiment “Fear” obtains the lowest subjective classification accuracy, this is perhaps because the acoustic characteristics of sentiment “fear” is not clear, this type of sentiments are mostly expressed through the linguistic information. The comparative performance analysis has been taken to evaluate the performance of proposed sentimental speech synthesis system illustrated in Table 2**.** Ten participators carefully listened the sentimental synthesized speech utterances, and recognized which type of sentiments they are.

Table 2. Comparative Performance Measure of Synthesized Speech and Natural Speech Output

Synthesized Sentimental Speech	**Natural Recorded Speech**			
	Angry	**Joy**	**Sad**	**Fear**
Angry	91.3%	8.7%	1.9%	0%
Joy	8.9%	90.1%	3.8%	2.1%
Sad	7.2%	7.3%	87.5%	7.2%
Fear	4.1	8.7%	12.1%	79.2%

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

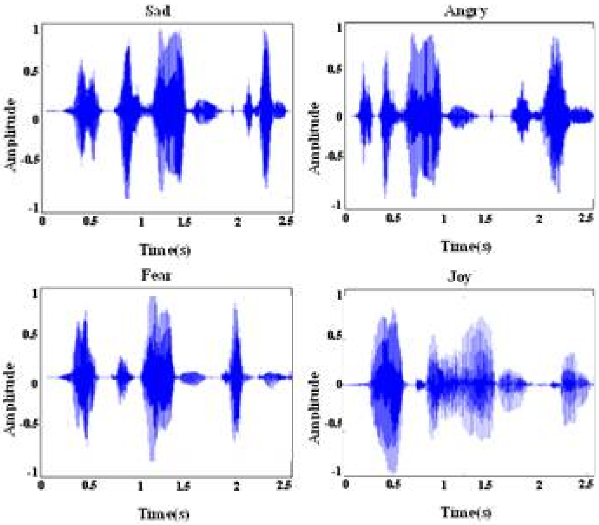

Figure 4. Sentimental Speech Output produced by the Vocal Track

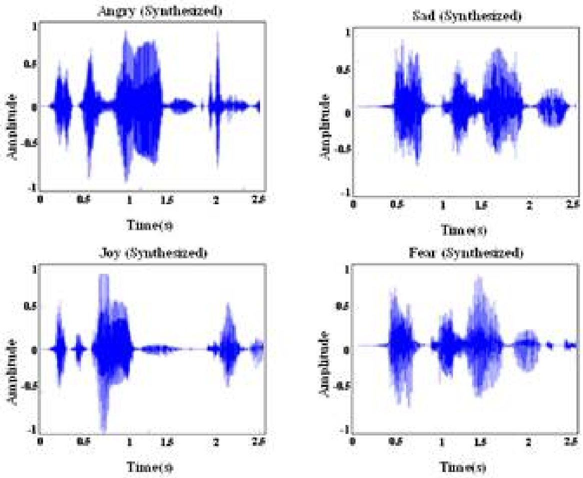

Figure5. Sentimental Synthesized Speech Output produced by the system

IX. CONCLUSION

In this proposed work a novel sentiment analysis are designed and implemented using prosodic feature modification method supported by TD-PSOLA technique. The major limitation of this work is that the sentimental speech data size is limited. To achieve an enhanced naturalness in the emotion expression, a bulk size of sentimental speech units database is needed. In order to generate an enhanced natural sentimental speech, the concatenative unit selected in this work is "word", because the speech database provides corresponding words for each type of sentiments. To achieve naturalness in the sentimental synthesized speech using smaller size of data set, have to be select shorter length of unit type, like syllables, phonemes, etc.

REFERENCES

1. Cahn, J. E., Generating Expression in SynthesizedSpeech, Master's Thesis, MIT, 1989.http://www.media.mit.edu/~cahn/masters- thesis.html
2. S.D. Shirbahadurkar, D.S. Bormane, R.L. Kazi. 2010. Subjective and spectrogram analysis of speech synthesizer for Marathi tts using concatenative synthesis, Recent Trends in Information, Telecommunication and Computing
3. M. Bulut, S. Narayan and A. Syrdal, "Expressive Speech Synthesis Using a Concatenative Synthesizer," Proceedings of ICSLP, 2002, pp. 1265-1268.
4. G. Hofer, K. Richmond and R. Clark, "Informed Blending of Databases for Emotional Speech Synthesis," Proceedings of Interspeech2005, pp. 501-504.
5. M. Schroder, "Speech and Emotion Research: An Overview of Research Frameworks and a Dimensional Approach to Emotional Speech Synthesis," Ph.D. Thesis, Saarland UniversitySaarland, 2004.
6. L. R. Rabiner and R. W. Schafer, "Digital Processing of SpeechSignals," Prentice-Hall, Inc., Englewood Cliffs,1978.
7. yang shun," Speech synthesis based on PSOLA algorithm and modified pitch parameters." Proceedings of Computational problem solving 2010, pp. 296-299.
8. D. Jurafsky and J. H. Martin. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2000
9. Fangzhong Su and Katja Markert. 2008. From word to sense: a case study of subjectivity recognition. In Proceedings of the 22nd International Conference on Computational Linguistics, Manchester
10. Burkhardt, F., Simulation emotionaler Sprechweise mit [Simulation of emotionalmanner of speech using speech synthesis techniques],PhD Thesis, TU Berlin, 2000.
11. Vroomen, J., Collier, R., & Mozziconacci, S. J. L.,Duration and Intonation in Emotional Speech,Eurospeech 93, Vol. 1, pp. 577-580.
12. F. Burkhardt, A. Paeschke, M. Rolfes, et al., "A Databaseof German Emotional Speech," Proceedings of Interspeech,2005.
13. S. Sangeetha, S. Jothilakshmi, "Syllable based text to speech synthesis system using Auto associative Neural Network prosody prediction", International Journal of Speech Technology, Springer publication, Vol. 17, no. 2, pp. 91- 98, 2014.