Telugu Emotional Story Speech Synthesis using SABLE Markup Language

Gurunath Reddy M, Harikrishna D M, K. Sreenivasa Rao and Manjunath K E
School of Information Technology
Indian Institute of Technology Kharagpur, India - 721302
mgurunathreddy@gmail.com, coolhari2006@gmail.com, ksrao@iitkgp.ac.in, ke.manjunath@gmail.com

***Abstract*—In this paper, a framework for synthesizing Telugu emotional speech for story telling applications is presented. An XML based markup langauge, SABLE is used to synthesize the emotions from a given story text. SABLE markup defines a set of tags to improve the quality of the synthesized speech from the concatinative speech synthesizer. In this work, a subset of prosody tags are used to synthesize the emotional speech from a given story text. Modified Zero frequency filtered (ZFF) signal is used to derive the prosody correlates of pitch base, pitch range and intensity. The desired prosody modification factors for each emotion is derived at phrase level. The derived prosody modification parameters for each emotion are stored in the form of a template. During synthesis, hand annotated story text is replaced by prosody tags which are stored in templates. Prosody tagged story text at phrase level is automatically converted into SABLE markup format. The markup story text is used to synthesize emotional speech from the Telugu neutral Festival TTS system. The quality and naturalness of the synthesised emotional story speech is evaluated using subjective tests.**

Index Terms— Emotions; ZFF; SABLE markup; Prosody tags; Emotional speech; Synthetic speech;

I. INTRODUCTION

Text to speech systems provides machines with the ability to speak. Speech synthesis technology has reached a stage where the synthesis systems provides natural and intelligible speech, but they lack in emotions (expressivity) [1]. Unit selection speech synthesis (USS) system is one such speech synthesis system which is capable of producing natural and intelligible speech for a given text input. In USS, the input text is converted to abstract linguistic representation by the front end Natural language processing (NLP) system. This linguistic representation is obtained by performing prosodic annotations on the syntactic, semantic and lexically analysed text [2]. This linguistic representation is used to pick the optimal units from the stored neutral spoken utterances. The optimal units are then concatenation to form the neutral synthesized speech utterance [3] [4] [5]. Such a system is called as neutral speech synthesis system. In emotional speech synthesis system, along with text, the desired emotion cue forms an additional input to the NLP. The input text is converted into abstract linguistic representation as in Neutral synthesis system. In addition, the emotional information is provided, either before or after the synthesis of neutral speech. In the former case, the emotional information is coded along with the linguistic information and speech is synthesized from the text using the linguistic and emotional information. In the later case, neutral speech is synthesized and then the desired expression is added using prosody or voice transformation techniques [6] [7].

From the existing literature, it is observed that there are some works related to emotional speech synthesis (ESS). Cahn (1989) developed the first ESS system on formant synthesizer: The Affect editor [8]. The control parameters of the synthesizer are manually tuned for each of the emo- tions to synthesize the emotive speech. Murray et al. (1995) developed Hamlet. Hamlet is a rule based ESS system built on commercial formant speech synthesis system, DECtalk. The pitch, duration and voice quality rules are defined in the DECtalk synthesizer for all emotions and quality of the synthesized emotions are improved by manual tuning [9]. Vroomen et al. (1993) synthesized emotional speech by ma- nipulating the pitch and duration of the neutral synthesized speech by pitch synchronous overlap add method [10]. Iida et al. (2000) demonstrated ESS by storing large databases for each emotion and target emotion is synthesized by selecting the units from respective emotional databases [11]. Campbell (2006) synthesized conversational speech by selecting phrasal units from a very large database [7]. Hofer et al. (2005) created a blend of database by mixing databases of angry, happy and neutral speech. The target emotional speech is achieved by penalizing the target cost for other than emotive units [12].

In the present work, the emotion specific parameters are obtained by analysing the prosody parameters of emotions. The prosody parameters includes pitch base, pitch range, intensity and speaking rate. The neutral synthetic speech and story emo- tions are processed by the ZFF method to estimate the prosody parameters based on the instantaneous pitch and strength of excitation. Phrase level prosody parameters are obtained and averaged for each emotion. The emotive speech is synthesized by annotating the input text with SABLE prosody tags (SPT). In this work, the text forms the linguistic information and SPT's forms the emotion cues for the synthesis engine. The SPT's guides the Festival Text to speech system (TTS) to select the optimal units based on the prosody factors provided in the prosody tags. The selected optimal units are concatenation to form the emotional speech.

Rest of the paper is organized as follows, Section II describes the story speech corpus collected for analysis. Proce- dure for deriving prosody parameters based on modified ZFF is discussed in Section III. Section IV describes story text tagging and synthesis of emotional speech from mark-up text. Evaluation and conclusions are discussed in Section V and Section VI respectively.

II. STORY SPEECH CORPUS

In this work, we have collected 60 Telugu children stories from various children story books. Stories are narrated by a female artist, having fair amount of experience in narrating children stories. A subset of 10 stories are selected for analysis from 60 children stories. The selected stories contained mainly Akbar-Birbal and Panchatantra stories. Akbar-Birbal and Panchatantra stories are chosen because, they contained fair amount of emotions. The emotion salient phrases are separated from the selected stories for analysis. The emotions considered for analysis are anger, happy, sad, fear. The total number of phrases for anger, happy, sad and fear are found to be 21, 26, 24 and 18 respectively. Neutral utterances are synthesized for all the phrases from Telugu neutral Festival TTS system. Hence, at the end, we had utterances with emotions and their corresponding neutral versions for analysis.

III. DERIVING PROSODY MODIFICATION PARAMETERS

A robust algorithm capable of capturing the rapid variations of source parameters is presented in the [13]. Hence, in this work, to capture the dynamically varying prosody parameters of emotional speech, the modified Zero frequency filtered based epoch extraction is used to derive the prosody parameters. The method uses zero frequency filtered signal derived from speech to obtain the instants of significant excitation and the strength of excitation at the epoch locations.

Zero-frequency filtered (ZFF) signal is derived as follows:

(a) The input speech signal s[n] is differenced to remove very low frequency components

$$x[n] = s[n] - s[n-1] \tag{1}$$

(b) The differenced speech signal is passed through a cascade of zero frequency resonators given by

$$y_0[n] = -\sum_{k=1}^{4} a_k y_0[n-k] + x[n] \tag{2}$$

where a1 = 4, a2 = 6, a3 = 4, and a4 = 1.

(c) The trend in $y_0[n]$ is removed by subtracting the mean computed over a window at each sample. The resulting signal y[n] is the ZFF signal, given by

$$y[n] = y_0[n] - \frac{1}{2N+1}\sum_{m=-N}^{N} y_0[n+m] \tag{3}$$

where $(2N+1)$ is the size of the window, which is in the range of 1 to 2 times the average pitch period in samples.

This method does not captures the rapid variations of F0 in the emotions. To capture the rapid variations of F0, the method is modified using the following steps to derive the epochs and their strength of excitation (SoE) from the ZFF signals.

(1) Speech signal is passed through the Zero-frequency resonator with a widow length of 3ms for trend removal. The energy of ZFF signal will be relatively high for voiced regions and low for unvoiced and silence regions.

(2) Hilbert envelope of the ZFF signal is computed. Envelope is smoothed by using a 10ms running average filter. Smoothed envelope is threshold by 1% of the maximum sample value to get voiced and unvoiced segments.

(3) After finding the voiced segments. For each voiced segment, window length for trend removal is obtained by location of maximum peak in the autocorrelation function of that segment. Each voiced segment is filtered separately with corresponding window length.

(4) The epoch locations are obtained from the positive zero crossings of the final filtered signal, strength of excitation (intensity) is obtained by the slope ZFF signal at epoch locations.

(5) Epoch interval plot is obtained by successive epoch location differences. Finally, pitch contour is obtained by the inverse of the epoch interval plot.

(6) The pitch and intensity contour plots for all neutral and emotion phrases are obtained. From the plots prosody parameters: pitch range (PR), pitch base (PB) and average intensity is calculated. The mean prosody parameters for all emotions: neutral, anger, fear, sad and happy phrases are obtained. The prosody modification factor is obtained by scaling the mean prosody parameter of the emotion with the mean neutral prosody parameter. For instance, the intensity modification factor of 30% to convert neutral to target angry emotion is obtained by taking the ratio of the average intensity of the angry to that of the neutral speech.

(7) Similarly the mean speaking rate modification factor for each emotion is obtained by measuring the mean rate of words for all phrases considered.

The pitch and strength of excitation contours obtained from the modified ZFF is shown in Fig. 1. Speech signal and the corresponding ZFF signal are plotted in Figs. 1(a) and 1(b). The voiced and unvoiced decision using smoothed threshold Hilbert envelope of ZFF signal is shown in Fig. 1(c). Fig. 1(d) shows the adaptive windowed ZFF signal. The Strength of excitation and pitch contours are shown in Figs. 1(e) and 1(f).

The prosody modification parameters derived at phrase level are shown in table I. The positive prosody modification factor in the table indicates that, the mean prosody parameter of corresponding neutral speech has to be increased to get the target emotion. Alternatively, a negative factor indicates that the mean prosody parameter of corresponding neutral speech has to be decreased to get the target emotion.

TABLE I. PROSODY PARAMETERS DERIVED FOR EACH EMOTION.

	Intensity(%)	Rate(%)	PR(%)	PB(%)
Angry	+30	+30	+112	+20
Happy	+20	+20	+150	+30
Sad	-10	-22	-30	-15
Fear	+10	-10	+43	+20
Neutral	+10	+10	+15	+20
Emphasis	+33	+10	+55	+20

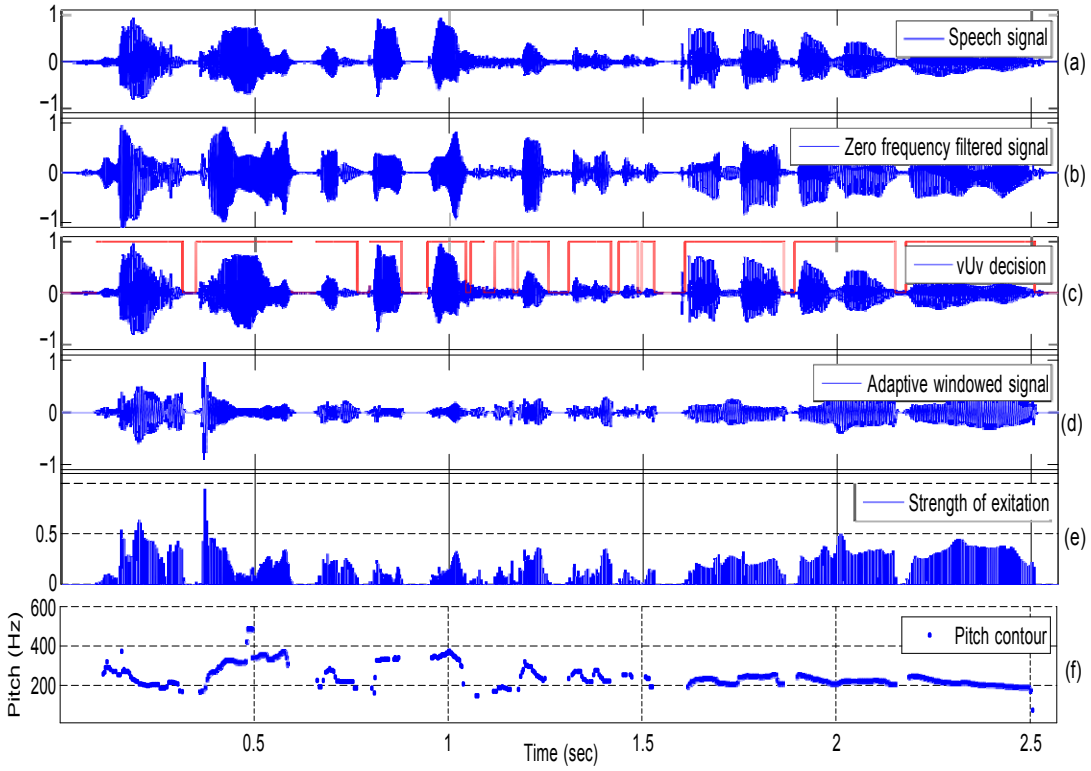

Fig. 1. (a) Happy emotive speech signal, (b) Zero frequency filtered speech, (c) voiced and unvoiced decision based on threshold of Hilbert Envelope, (d) Adaptive windowed ZFF signal to capture the dynamic varying pitch in happy emotion utterance, (e) strength of excitation signal, (f) pitch contour obtained from epoch interval plot.

IV. Story Text Tagging and Synthesizing Emotional Speech

In this work, SABLE mark-up [14] is used for synthesizing the emotional speech. SABLE is used because it defines a variety of XML tags for increasing the quality of synthesized speech. Also, our neutral Telugu TTS system is built upon Festival framework and SABLE is well supported by Festival. Here, we have used a subset of tags from the SABLE which modifies the vocal features of the neutral speaking TTS system. The prosody tags used for emotional speech synthesis are as follows:

Pitch base, defines the baseline pitch offset of TTS voice. It is denoted by the tags:

<PITCH BASE="± modification factor%">
</PITCH>

Pitch Range, defines the vocal range offset, denoted by the tags:

<PITCH RANGE="± modification factor%">
</PITCH>

Rate of speech controls the speed at which the text is spoken, represented by the tags:

<RATE SPEED="± modification factor%">
</RATE>

Volume level controls the intensity of speakers voice, and is denoted by the tags:

<VOLUME LEVEL="± modification factor%">
</VOLUME>

Emotional story synthesis using SABLE mark-up language involves identifying emotional salient phrases from the story text and annotating each phrase with simple emotion tags. Further, annotated text is converted into SABLE mark-up format using emotion templates and story is synthesised using Festival TTS system. An indication must be given to the synthesis system to synthesize the appropriate emotion for a given phrase. Hence, the input story files are hand annotated with simple emotion tags. The simple emotion tags includes: neutral, angry, happy, sad, fear and emph (stands for emphasis). Story text is manually tagged using the tags of the form [EMOTION] at the beginning of each emotion phrase. The tag, EMOTION is an element of the set: anger, happy, sad, fear, neutral and emph. An example of input Telugu story text manually tagged with simple emotion tags is shown in Figure 2. The example story is about a proud deer for its beautiful long legs and horns. On an unfortunate day, lion attacks the deer while drinking water in a small pond. The deer fails to escape form the lion due to its log legs and horns.

Emotion tagged phrases in the story text need to be synthesized with the corresponding emotions. The emotions are characterised mostly by the prosody features. The procedure for deriving the prosody modification factors for converting

neutral to target emotion are presented in Section III. The prosody modification factors for five emotions angry, happy, sad, fear and neutral are tabulated in Table I. From the table it can be observed that angry and happy emotions have the largest average pitch range, intensity and speaking rate compared to all other emotions. This because a person in happy or anger state try to speak faster with louder voice, which results in faster vocal folds vibration than normal and higher strength of excitation. Sad has the lowest average pitch, intensity and speaking rate because, a person in sad speaks with very low intensity and speaking rate which results in slower vibration of vocal folds and lowered strength of excitation. Here, we have considered neutral is also a emotion because the speaking rate, intensity and pitch range of the neutral utterances of story narrator differed from the synthesized neutral utterances. From the table it can be observed that the speaking rate and intensity of the neutral utterances of the narrator is 10% higher than the synthesized neutral speech. Pitch range and pitch base are found to be 15 and 20% higher then synthetic neutral speech. While analysing stories, it is observed that, most of the time the narrator emphasised some words to keep the attention of the listeners. The emphasised words are mostly adjectives. Separate prosody modification factors are derived for emphasis words and are shown in table I.

[neutral] ఒక రోజు అందమైన కొమ్ముల జింక ఒకటి , [neutral]నీటి మడుగులో నీరు తాగుతుంది. [neutral]నీటిలో దాని ప్రతిబింబం కనిపిస్తోంది. [neutral]నీటిలో కనిపిస్తున్న తన కొమ్ములను చూసుకొని, [emph]ఆహా! [happy]ఎంత అందమైన కొమ్ములు, [neutral]అని తన్మయంగా అనుకుంది. [neutral]అది ఇంకొంచెం ముందుకు నడిచింది. [neutral] ఈ సారి దాని ముందర కాళ్ళు నీటిలో కనిపించాయి. [emph]బలిష్టంగా [emph]పొడవుగా [neutral]ఉన్న కాళ్ళను చూసి , [emph]చ! [sad]ఇవి వికారంగా ఉన్నాయి అని అనుకుంది జింక. [neutral]అంతలో దానికి సింహం గర్జన వినిపించింది, [neutral]వెను తిరిగి చూసింది. [neutral]తనపై దాడి చేయడానికి సింహం వస్తుండడంతో పరుగు తీసింది . [sad]అయితే దాని [emph]బలిష్టమైన, [emph]పొడవైన [neutral] కాళ్ళు సింహం నుండి తప్పించలేక పోయాయి. [sad]దురదృష్టవశాత్తు దాని కొమ్ములు, [neutral]ఒక తీగలో చిక్కుకున్నాయి. [sad]అది ఆ తీగ నుండి విడిపించుకోలేక, [neutral]సింహానికి దొరికిపోయింది.

Fig. 2. Sample text story annotated with simple tags.

The prosody parameters derived for all the emotions are stored in the form of a emotion template. The emotion template is as shown in the Figure3. From the figure it can be observed that each template consists of emotion name and three prosody tags: pitch, volume and intensity. The prosody modification factor for the emotion is specified in the corresponding tags. During synthesis, the simple annotated emotions tags in the story text are automatically parsed and the text phrases are embedded into the corresponding emotion templates. The embedded phrase templates are automatically converted into SABLE mark-up format where the Festival can understand the prosody tagged sentences to synthesise the emotional speech. Two phrases from the Figure 2 converted to SABLE mark-up format is shown in Figure 4.

```
<emotion name="happy">
        <PITCH RANGE="+50%" BASE="+30%">
        <VOLUME LEVEL="+20%"/>
        <RATE SPEED="+20%"/>
</emotion>
<emotion name="angry">
        <PITCH RANGE="+100%" BASE="+20%"/>
        <VOLUME LEVEL="+30%"/>
        <RATE SPEED="+30%"/>
</emotion>
<emotion name="sad">
        <PITCH RANGE="-30%" BASE="+30%"/>
        <VOLUME LEVEL="-10%"/>
        <RATE SPEED="-20%"/>
</emotion>
<emotion name="neutral">
        <PITCH RANGE="+15%" BASE="0%"/>
        <VOLUME LEVEL="0%"/>
        <RATE SPEED="+10%"/>
<emotion name="fear">
        <PITCH RANGE="+40%" BASE="0%"/>
        <VOLUME LEVEL="0%"/>
        <RATE SPEED="+10%"/>
</emotion>
```

Fig. 3. Emotion templates for happy, angry, sad, fear and neutral emotions.

```
<SABLE>

<!-- neutral -->
 <PITCH RANGE="+15%" BASE="0%" >
 <VOLUME LEVEL="0%">
 <RATE SPEED="+10%">
   ఒక రోజు అందమైన కొమ్ముల జింక ఒకటి ,
 </RATE>
 </VOLUME>
  </PITCH>
</SPEAKER>

<!-- neutral-->
 <PITCH RANGE="+15%" BASE="0%" >
 <VOLUME LEVEL="0%">
 <RATE SPEED="+10%">
   నీటి మడుగులో నీరు తాగుతుంది.
</RATE>
</VOLUME>
 </PITCH>
</SPEAKER>
</SABLE>
```

Fig. 4. Two text phrases from input story converted to SABLE mark-up format.

V. EVALUATION

Traditionally emotional speech is evaluated by presenting a number of synthesised emotional utterances to the subjects. A forced choice is made by the subjects to choose emotion categories based on the perceived prosody of the test utterance [15] [16]. In the forced choice experiments, subjects lack the additional signs present in the natural situations. Hence, it is argued that perception is not very accurate in this type of experiments [17]. For applications like story speech synthesis, whether the prosody fits with the message is more important than subjects can classify the prosody without additional signs or cues [18].

Therefore, following [18] [19] proposed prosody rules are evaluated in a naturalistic context, using utterances which are clearly identifiable as fragments of children stories. As material, we have used 16 short utterances from four stories, taken from the stories used for analysis. Two categories of utterance are created for each test fragment. One category generated by neutral Telugu TTS system without prosody modification and the second category generated by using proposed prosody modification factors. Five subjects participated in the evaluation test. The test fragments are presented to them in a randomized order. After a short introduction to the experiment, the subjects listened to a short synthesized speech utterance, intended to make the subjects comfortable with synthetic speech. After that each subject presented with the 32 synthesised utterances. For each utterance, the subjects are asked to rate its storytelling quality with embedded emotions and naturalness on a five point scale (1 = very bad, 5 = excellent). The subjects are also asked to provide free comments about each utterance. Table II shows the results of subjective evaluation. The average ratings for the storytelling and neutral utterances are given in A and B columns respectively. From the table, it can be observed that the prosody modified utterances judged to have higher storytelling style than the prosodically neutral versions with an average score of 3.36. Furthermore, the storytelling utterances are judged more emotional than the neutral utterances. From the naturalness evaluation column we can observe that the naturalness of few storytelling utterances judged to be slightly lower than that of the neutral versions. Relatively higher prosody modification factors used for anger and happy emotions resulted in the lower ratings for naturalness, which introduced audible glitches in the generated speech.

TABLE II. EVALUATION RESULTS(A = STORY STYLE B = NEUTRAL)

	Story Telling		Naturlness	
Utterance	A	B	A	B
1	3.4	3.0	3.3	3.1
2	3.9	3.1	3.3	2.9
3	3.5	3.2	2.9	3.0
4	4.1	2.6	3.5	3.2
5	3.3	3.1	3.1	3.2
6	3.3	2.9	3.4	3.1
7	2.9	3.0	3.2	2.5
8	3.5	3.2	3.4	2.6
9	3.1	3.1	3.9	2.6
10	3.4	2.5	3.5	3.1
11	3.2	2.6	4.1	2.9
12	2.8	3.2	3.3	3.0
13	3.1	3.1	3.5	3.2
14	3.5	2.9	3.1	2.5
15	3.2	3.1	3.4	2.6
16	3.6	3.2	3.2	3.2
Total Avg.	3.36	2.98	3.38	2.91

VI. SUMMARY AND CONCLUSIONS

A preliminary work is presented for synthesizing emotional speech for story telling application in Indian language: Telugu. A set of prosody modification factors are derived at phrase level and used for synthesizing story speech. The presented work is evaluated using subjective tests. Subjects rated the synthesised speech based on storytelling quality and naturalness. From the results, it is observed that, the synthesized stories are indeed perceived as natural and story telling.

ACKNOWLEDGMENT

This work is carried out at IIT Kharagpur as a part of project titled "Development of text to speech (TTS) synthesis system for Indian languages (Phase-II)" supported by Department of Information Technology, Govt. of India.

REFERENCES

[1] J. Y. Zhang, A. W. Black and R. Sproat, "Identifying Speakers in Children's Stories for Speech Synthesis," in *INTERSPEECH, Geneva*, 2003.

[2] D. H. Klatt, "Review of text to speech conversion for English," *The Journal of the Acoustical Society of America*, 1987.

[3] D. H. Klatt, "Software for a cascade/parallel formant synthesizer," *The Journal of the Acoustical Society of America*, 1980.

[4] N. P. Narendra and K. Sreenivasa Rao, "Optimal weight tuning method for unit selection cost functions in syllable based text-to-speech synthesis ," *Applied Soft Computing (Elsevier)*, 2013.

[5] N. P Narendra and K. Sreenivasa Rao, "Syllable Specific Unit Selection Cost Functions for Text-to-Speech Syntheisis," *ACM Transactions on speech and language processing*, 2012.

[6] J. Tao, Y. Kang, and A. Li, "Prosody conversion from neutral speech to emotional speech," *IEEE Transactions on Audio, Speech, and Language Processing*, 2006.

[7] N. Campbell, "Conversational speech synthesis and the need for some laughter," *IEEE Transactions on Audio, Speech, and Language Processing*, 2006.

[8] J. E. Cahn, "Generation of affect in synthesized speech," *In Proceedings of American voice I/O society*, 1989.

[9] I. R. Murray and J. L. Arnott , "Implementation and testing of a system for producing emotion by rule in synthetic speech," *Speech Communication*, 1995.

[10] J. Vroomen, R. Collier and S. J. L. Mozziconacci, "Duration and intonation in emotional speech," *In the Proceedings of EUROSPEECH*, 1993.

[11] A. Iida, N. Campbell, S. Iga, F. Higuchi and M. Yasumura, "A speech synthesis system for assisting communications," in *In ISCA workshop on speech and emotion*, 2000.

[12] G. Hofer, K. Richmond and R. Clark, "Informed blending of databases for emotional speech synthesis," in *In the Proceedings of INTERSPEECH*, 2005.

[13] K. Sudheer, M. S. H Reddy, K. S. R. Murty and B. Yegnanarayana, "Analysis of laugh signals for detecting in continuous speech," in *Proceedings of Interspeech, Brighton, UK*, 2009.

[14] R. Sproat, A. Hunt, M. Ostendorf, P. Taylor, A. Black, K. Lenzo and M. Edgington, "SABLE: A standard for TTS markup," in *International Conference on Spoken Language Processing, Sydney, Australia*, 1998.

[15] J. Cahn, "The generation of affect in synthesized speech," *J. American Voice I/O Society*, vol. 8, 1990.

[16] M. Bulut S. Narayanan and A. Syrdal, "Expressive speech synthesis using a concatenative synthesizer," in *Proceedings of International Conference on Spoken LanguageProcessing*, 2002.

[17] L. Van Santen, G. Black, A. Cohen, E. Kain, T. Klabbers, J. Mishra de Villiers and X. Niu, "Applications of computer generated expressive speech for communication disorders," in *Proceedings of Eurospeech, Geneva, Switzerland*, 2003.

[18] M. Schroder, "Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions," in *Proceeding Workshop on Affective Dialogue Systems, Kloster Irsee, Germany*, 2004.

[19] W. L. Johnson, S. Narayanan, R. Whitney, R. Das, M. Bulut and C. LaBore, "Limited domain synthesis of expressive military speech for animated characters," in *IEEE Speech Synthesis Workshop, Santa Monica, USA*, 2002.