International Journal of Enhanced Research in Science Technology & Engineering, ISSN: 2319-7463
Vol. 3 Issue 2, February-2014, pp: (348-352), Impact Factor: 1.252, Available online at: www.erpublications.com

Text-To-Speech Synthesis for Marathi and Prosody Conversion

Rohit S. Deo[1], Pallavi S. Deshpande[2]
Dept. of E&TC, SKN College of Engineering, Pune, India

Abstract: Text-To-Speech synthesizer converts the input text into the speech waveform. The generated speech from such a Text-To-Speech synthesis system is generally emotionless and monotonous. So as not to get it to the listener boring, it is needed to add prosody into the synthetic speech. The objective of the proposed work is to modify the prosody of a neutral speech, generated by a Marathi text-to-speech synthesizer. The synthesizing system works on unit selection principle. The present paper sheds light on the prevalent approaches of speech generation and adding expressivity by modifying the neutral speech parameters. Unit selection based system generates the synthetic neutral speech and its spectral parameter variation pours expressivity to it. From the speech generated from TTS system, stressed word and syllables are identified to add expressivity in the form of interrogate. Replacing the pitch tier the result is obtained. Results of the subjective experiments show the effectiveness of the system.

Keywords: Hidden Markov Models, Classification and regression tree, Text-to-speech.

I. INTRODUCTION

Speech synthesis is the process of conversion of written text into a speech signal. Expressive speech synthesis comprises of adding expressivity to the neutral speech. The state of the vocal tract of human being varies as the emotion varies. This variation is reflected on speech waveform, called as an expression. Synthesized speech with different expressions can be used as a story telling application for children. The monotony in the speech generating systems makes the listener boring. Adding prosodies with appropriateness is a bit crucial task. Adding expressivity to a synthetic voice has two aspects: being able to realize an expressive effect, i.e. having the technological means to control the acoustic realization; and having a model of how the system's realization should sound in order to convey the intended expression. Expressive speech synthesis (ESS) involves adding prosodies to the monotonous synthesized speech. This can be done by controlling expressivity by training statistical models based on different databases.

The fundamental frequency (also called the fundamental) of a periodic signal is the inverse (reciprocal) of the pitch period length. The F0 frequency is a measure of how high or low the frequency of a person's voice sounds. Its psychological correlate is pitch. It is the frequency of vocal fold vibration and correlates with changes in vocal fold tension and sub-glottal air pressure. The nature of pith contour and duration pattern (for e.g.- duration of a syllable, number of pauses, position pauses, steepness of f0 contours during rises and falls) highly represent the emotion. There are many parameters such as jitter, shimmer, Mel-frequency cepstral coefficients, kurtosis, intensity etc. that can be used for modelling prosody. The explicit way to control the parameter and its nature is to use copy synthesis wherein the control is made by using natural rendition of emotion. Instead of recording every diphone with neutral pitch, different versions of diphone units are recorded in order to make the monotonous speech human-like. So, ESS involves the feature extraction and parameterization of target expression. Feature extraction characterizes an object whose values are similar for the same category and different for different categories by measurement. So, distinguishing features that are invariant must be extracted by an ideal feature extractor followed by parameterization. Feature extraction must seek the distinguishable features that are invariant to the irrelevant transformation. Assimilating all the features, their uniqueness, optimizing them, a model is generated for the generation of speech with expressivity. The generated model is then trained with the sample database for efficient synthesis of speech. The rest of the paper is organized as follows. Section 2 gives literature survey. Section 3 elaborates proposed unit selection based TTS system. Modelling and modification of neutral speech is described in section 4 followed by results and discussions in section 5. Section 6 gives conclusion.

II. LITERATURE SURVEY

Rule based formant synthesis, one of the oldest speech synthesis techniques, estimates the acoustic realization of speech waveform based on the constraints, defined by expert hands. Later, Concatenative speech synthesis evolved. The system chopped human voice recordings at diphone level and while synthesizing speech, re-sequenced it. Different pitch contours and duration patterns can be realized using the techniques such as pitch synchronous overlap-add

(PSOLA). Oytun Turk and Mark Schroder determined the capability of the voice transformation for the speech synthesis targets also the importance of the intonation contour as compared to the f0 contours, its range and level [1]. GMM is utilized for the transformation of voice quality into expressive targets e.g. - neutral speech to desired emotional prosody such as depressed, cheerful or aggressive. Effective uses of frequency domain pitch synchronous overlap add (FD-PSOLA) has kept the ball of progress for expressive speech synthesis of neutral speech rolling [2]. One of the popular methods for the expressive speech generation is based on Hidden Markov Models. Here, description of the algorithm for the generation of speech parameter sequence; emerged from the spectral analysis of the speech is done [3][4]. Effective use of multi-mixture HMMs produce clear formant structure as compared to that of single-mixture HMMs. Based on dynamic feature parameter vector & spectral parameter vectors, feature vector is created. Jianhua Tao, Yongguo Kang, and Aijun Li converted prosody into an emotional speech by classifying the speech into strong, medium and week. It used three different models for experimentation viz. – classification and regression tree (CART), Gaussian mixture model (GMM), [5] and Linear modification model. CART maps the integrated linguistic features whereas LMM directly operates on F0 contours of the speech signal. By modifying different dimensions of F0 contour such as F0 duration, intensity, steepness of rise and fall, number of pauses and duration of pause etc.

Modelling the pitch and state duration is accomplished in [6]. As a spectral parameter they used MFCC whose sequence vector is obtained from speech database. MSD-HMM based pitch patterns are poured into the system. To avert the discrepancy amongst pitch model and spectral model, fabric of context dependent MSD-HMM is made, trained with features. In extension to [6], [7] changed voiced characteristics of speech, synthesized for speech recognition by using technique of speaker adaptation at ease. The synthesized speech has the vocoder type of quality which can be improved further by post-filtering. By 2012, Jing Zhu and Yibiao Yu used intonation and prosody for the expressive speech in Mandarin. While studying the prosody pattern, the speaking speed, pause and accent of the speaker is considered. As per the requirement, respective patterns of prosody are modified. To make it more expressive, intonation along with prosody is used. To describe the voice pathologically, the two parameters viz. – Jitter and Shimmer are widely used [8]. Cycle to cycle variations of F0 and amplitude are measured in terms of jitter and shimmer. Experiments show that measuring jitter and shimmer give marvellous results as feature vectors of prosodic and spectral parameters [9]. They provide complementary information. In more recent studies, use of articulatory features‟ use in expressive speech synthesis is elaborated. It is showed that, articulatory features can be predicted from text and human-like or naturalness in speech can be constructed. Both the subjective and objective testing gave good results.

III. TEXT TO SPEECH SYNTHESYS FOR MARATHI

To make synthetic speech human-like, prosody of neutral speech is modified by studying the acoustic and statistical models of target expression. The generation of synthetic speech is made by text-to-speech systems. A large database is recorded of a speaker with clear and consistent voices. The purpose of this work is to generate expressive speech from neutral speech outputs of a Marathi TTS synthesizer. The basic unit selection premise is that we can synthesize new naturally sounding utterances by selecting appropriate sub-word units from a database of natural speech.

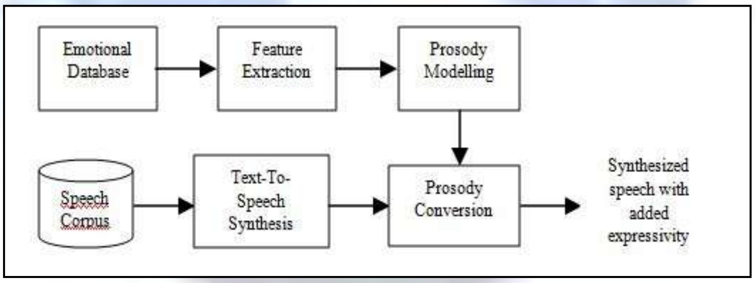


Fig. 1: Block diagram for expressive speech synthesis for Marathi

A. Recording of database

A text prompt of 1000 sentences in Marathi is derived which contains 10,000 words and syllables. The same text is recorded by a native Marathi speaker, having several years of research experience in speech processing in a professional studio. Such .wav files with 16-bit PCM coding with mono-channel and 16 KHz sampling rate is stored as database. This data is needed for data analysis and modeling.

B. Declaration of phoneset

The lingual characteristics of Marathi are created in phoneset along with its phoneme features. The feature values are added to make it more appropriate for Marathi. The phone features such as vowel or consonant, its length (short, long, diphthong, schwa) are taken into consideration. Also, height of vowel (high, mid, low), frontness of vowel (front, mid back), and necessity of lip rounding is analysed and added. There are different types of consonants in Marathi such as fricative, affricative, nasal, liquid. The feature values for each consonant are set with appropriate place articulation. It may be labial, alveolar, palatal, labiodentals etc. By deliberating on all these constraints, a feature vector is created.

C. Generation of prompts

A parser is needed to be declared with help of the declared phoneset. Once all this setup is ready, prompts are generated. For each sentence (i.e. - for text and its respective wav file), a prompt comprised of phones, words, syllables, F0 frequency, phoneme duration is made. Then, to align speech utterances with the text, auto labelling is done. Extracting and marking of pitch marks is done to label the utterances. Then pitch markers and Mel-Cepstral features are converted into labelling.

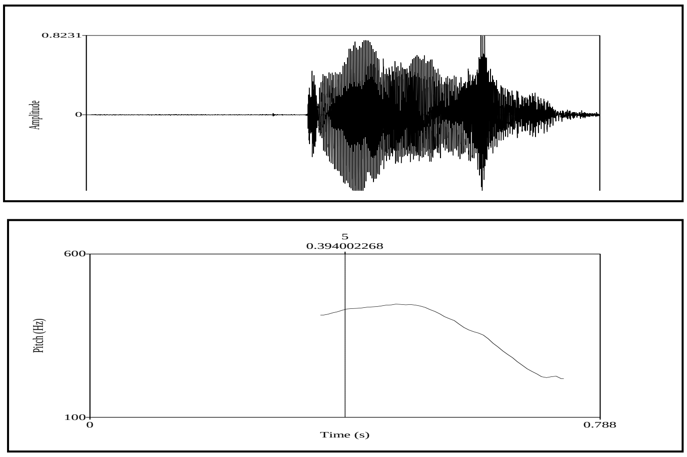
D. Clustering

Clustering of all the text prompts, its chopped sound wav file, its phoneset, labeling is done to generate a catalogue and a tree. A Classification and regression tree (CART) algorithm is used to synthesize a Marathi spoken sentence. For the written text in Marathi, with the use of scripting used to create text prompts, output sound file is generated in the speakers' voice. When this Text-To-Speech synthesizer is given an input in the form of text in Marathi scripts, it starts its execution. The entered scripted text is normalized by removing all non-sounding symbols from it. This helps in building tokenizer for the script. Tokenizer chops a scripted sentence into words and removes silence (if any) is there in the database. Pre-defined parser chops each word segment of the input till the phoneme level is reached. i.e. – it converts words into phoneme units. A classification and regression tree (CART) algorithm plays vital role. The decision tree generates exact sequence by hashing. And alas, a synthesized speech sound file in wav format is generated.

IV. CONVERSION OF PROSODY

The output of Marathi TTS synthesizer is used to modify the prosody. The variation of acoustic features of neutral speech may lead it to convert into expressive speech. Using F0 contours and various duration patterns, obtained at the cost of increased distortion, the features are fixed. These features are nothing but the acoustic correlates of a target expression. This gives better results than the formant synthesis. By cross combining different voice qualities and expressive styles, it is found that for a given speaker, the relative contribution of prosody and voice quality depends on the expressed emotion. Feature extraction must seek the distinguishable features that are invariant to the irrelevant transformation.

Psychoacoustic experiments have shown that F0 contour contains prosody information in it and it varies from emotion to emotion. Hence, F0 contour and duration are widely used to change the emotion in the synthetic speech. The difference in the underlying relations between target and source curves is explored. There are different approaches that change the pattern of F0 contour. Guassian normalization is used for mapping the values from reference to desired pitch values. Discrete Time Warping (DTW) is also used. The proposed work replaces the pitch tier of neutral sound wav by the target emotion pitch contour preceded by modification in duration tier of neutral word according to the target duration tier. Fig. 2 illustrates the waveforms of neutral and desired emotion along with their respective pitch contours. It can be easily seen that, the nature of pitch contour is much diversified as emotion varies. While uttering a sentence, it is observed that, prosody never prevail the whole sentence. Instead, speaker usually emphasizes on a particular word and syllable. Hence, while modifying expressions from neutral, only stressed syllable is under test. Such syllabic words are identified; their respective target expression is recorded. The proposed work modifies a neutral Marathi word into interrogative expression. F0 contour and duration pattern play vital role in modifying the prosody.



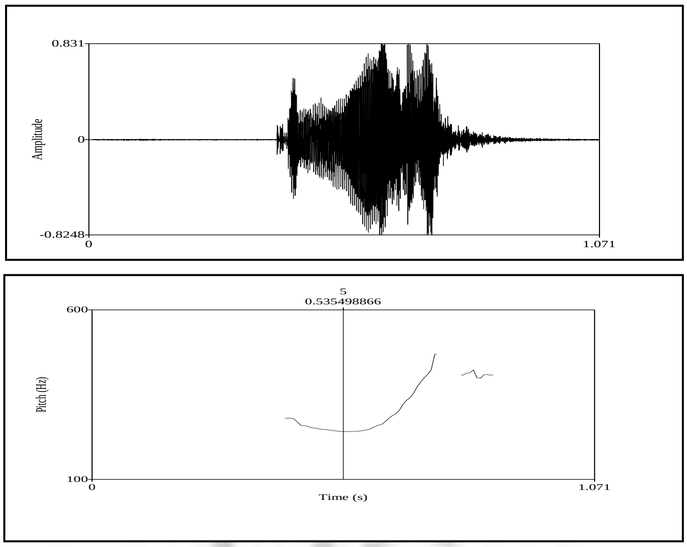


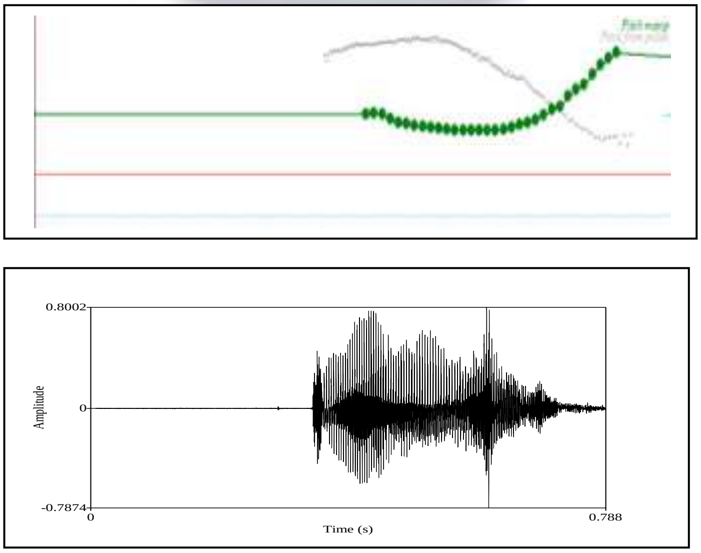
Fig. 2. (a) speech waveform of neutral Marathi word “kalala” (b) neutral waveform’s pitch contour (c) Target emotion (interrogative) waveform (d) its pitch contour

V. RESULTS AND DISCUSSIONS

As described in fig. 3, pitch contour of interrogative word is manipulated on the neutral Marathi word “kalala” to get it converted into interrogative “kalala”. The nature of the pitch contour of neutral sound wave is much different from the nature of the pitch contour of the target interrogative emotion. After replacing the pitch tier of neutral sound by the one of target, obtained resynthesized waveform and its pitch contour is shown in it. As per the observation that, emotion in a sentence is prevailed in stressed word and syllables, such Marathi stressed words in sentences are selected. Selection was done by five common people. Such ten words were under experiment and their pitch and duration tiers were modified according to the desired interrogative emotion. The experiment was subjected to subjective test of those five common people who selected the words. Each one of them was asked to score it on the basis of a scale, defined below. Table I illustrates, how the subjective analysis has been scored. The average score got from them is four. The resynthesized speech waveform indulged them after adding prosody.

TABLE I.

Score	Opinion	% Indulgence
1	Badly matched	< 20%
2	Unnatural	< 40%
3	Acceptable	Up to 50%
4	Natural	< 75%
5	Expressive	Up to 100%



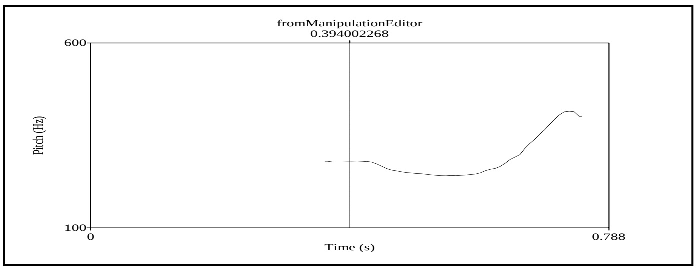


Fig. 3. (a) Replacement of pitch tier (b) Resulting speech waveform of "kalala" with addedprosody (c) its pitch contour

VI. CONCLUSIONS AND FUTURE SCOPE

To add naturalness to the monotonous speech, expressive speech synthesis points out concatenated and statistical models, making speech more human like. The proposed work frames out the extraction of features from speech and synthesize it. Unit selection based Text-To-Speech synthesis system works to the best and it makes generated speech more human like. This generates neutral speech. The synthesized output speech waveform from the Text-To-Speech synthesis system indulges the native listeners. Adding prosody of interrogate in the stressed syllable modifies the emotion of the neutral speech word. The nature of pitch contour and duration varies from emotion to emotion. It is the base to convert the prosody of neutral speech. The way to modify the prosody in the present work is proven to be the most convenient and easiest. In near future, with help of same technique of modifying prosody, more emotions are supposed to be added such as angry, joy, good news, bad news, confusion, apology, confidence etc. Also, Text-To-Speech synthesis system for Hindi.

REFERENCES

[1]. Oytun Turk and Mark Schroder, "Evaluation of expressive speech synthesis with voice conversion and copy resynthesis techniques", IEEE trans. on audio, speech and language processing , vol. 18., no. 5, July 2010, pp. 965-968

[2]. Jerome R. Bellegarda, "A data driven affective analysis framework towards naturally expressive speech synthesis" IEEE trans. on audio, speech and language processing , vol. 19, no. 5, July 2011, pp. 1113-1115

[3]. Yixiong Pan, Peipei Shen, Liping Shen, "Feature Extraction and selection in speech emotion recognition" National natural science foundation of China under Grant No. 60873132, pp. 64-68.

[4]. Keiichi Tokuda, Takayoshi Yoshimura, Takashi Masuko, Takao Kobayashi, Tadashi Kitamura, "Speech parameter generation algorithm s for HMM-based speech synthesis", Ministry of Education, Science, Sports and Culture, Japan, Grant-in-aid for Scientific Research (B) 2, 1055125, 1998, Encouragement of Young Scientists (A), 10780226, 1998.

[5]. Jianhua Tao, Yongguo Kang, and Aijun Li, "Prosody conversion from neutral speech to emotional speech" IEEE trans. on audio, speech and language processing , vol. 14., no. 4, July 2006, pp. 1145-1148

[6]. Takayoshi Yoshimuray, Keiichi Tokuday, Takashi Masukoyy, Takao Kobayashiyy and Tadashi Kitamura, "Simultaneous modelling of spectrum, pitch and duration in HMM-based speech synthesis", Ministry of Education, Science, Sports and Culture, Japan, Grant-in-aid for Scientific Research (B) 2, 1055125, 1998, Encouragement of Young Scientists, 0780226, 1998.

[7]. Keiichi Tokuda, Heiga Zen, Alan W. Black, "An HMM-based speech synthesis system applied to english", at Carnegie Mellon University.

[8]. Jing Zhu and Yibiao Yu, "Intonation and prosody conversion for expressive mandarin speech synthesis", ICSP 2012 proceedings, 978-1-4673-2197-6, pp. 549-551.

[9]. Mireia Farrús, Javier Hernando, Pascual Ejarque, "Jitter and shimmer measurements for speaker recognition".