Neutral to Emotional Speech Conversion by Pitch Contour Modification for Marathi

Rohit S. Deo
Dept. Of E&TC
SKN College of Engineering,
Pune, India

Pallavi S. Deshpande
Dept. Of E&TC
SKN College of Engineering,
Pune, India

***Abstract*— In this paper, we have used pitch contour models for modifying prosody of the neutral speech, generated from a Text- To-Speech synthesis system for Marathi. A pitch target model is applied to model and modify the prosody of the Marathi words in the form of interrogate. The proposed approach starts with the existing pitch contour of words that are needed to be stressed while adding expressivity into it. The proposed algorithms initiate with an existing approach of Gaussian Normalization for pitch mapping. Later, scatter plot pitch modeling and nth order polynomials are used to model and modify the prosody. Results of the subjective experiments show the effectiveness of the system.**

***Keywords*— Text-to-speech, Expressive Speech Synthesis, Prosody**

I. INTRODUCTION

Amongst existing speech generating systems, appropriate rendering of speech is unreachable. The monotony in the speech generating systems makes the listener boring. The work has been carried towards adding expressivity to the synthetic speech so as to make it more interesting to the listener. Since 1980, the work in expressive speech synthesis is carried. As the time passed, different techniques evolved with its strengths and demerits. Expressive speech synthesis comprises of adding expressivity to the synthetic voice such as happy or subdued, friendly or empathic, authoritative or uncertain. Adding prosodies with appropriateness is a bit crucial task. Analyzing different databases and studying its characteristics, modeling its nature of curves in statistical and mathematical sense is goaled.

It is observed that; pitch pattern variation represent the intonation of words to be uttered. In normal speech, pitch contour has emotional and non-linguistic type of information which play a vital role. By dynamically combining different factors such as muscle tension of larynx, elasticity and length of vocal tract, and sub-glottal air-pressure vocal folds slowly oscillate to realize a pitch contour. Emotional modulation affects the important properties of pitch contour. In Marathi, intonation diversifies as a function of stress and can convey emotions such as interrogate. Further the same model can be used to model and modify other emotions. Hence, pitch contour prevail the world of modifying prosody. Accordingly, benefit lies in modeling this dynamicity of F0 contour of human speech naturalness. Many times conversion of f0 is simply done through mean-variance transformation. Prosodic parameters such as f0 level, f0 mean, etc. are used to classify the emotional speech. An accurate and appropriate structure of pitch contour is important. There are different approaches to generate pitch contour. Both the time and frequency domain properties are used for pitch detection and finally contour. Use of higher level statistical information such as mean, range, minimum, maximum of f0 and energy helps to create a feature vector. Many a times neutral speech models are build to discriminate it from emotional speech.

The paper sheds light on modifying the pitch contour of stressed words which manifest interrogative emotion in Marathi. We start with the target pitch contour, its modeling and modification of surface pitch contour in accordance with the target one. The rest of the paper is organized as follows. Section 2 gives literature survey. Section 3 elaborates Gaussian normalization mapping of pitch contours. Modeling and modification of neutral speech and target by using nth order polynomials is described in section 4 followed by results and discussions in section 5. Section 6 gives conclusion.

II. LITERATURE SURVEY

Jan P. H. Van Santen in his research the issue of how to describe a pitch contours. Precise timing of local pitch accent curves in association with accented syllables has vital effects on how the interpretation of utterance is done by listeners. He showed that though small changes in alignment are audible, utterances alter the intentional meaning [1].

A framework is done in [2] for accounting variations in pitch contours. The pitch targets are defined and rules for their implementations are framed into it. The implementation rules are based on possible articulatory constraints on the production of surface F0 contours. Pitch targets are defined as the smallest operable units associated with linguistically functional units. Targets are either static or dynamic. The constraints on targets results in the partial reflection of the underlying pitch targets for implementing simple pitch target. The discussion about the observation of F0 patterns, with carryover and anticipatory variations, F0 peak alignment, declination and downstep. Yongguo kang and Jianhua tao employed the pitch target model for the representation and conversion of F0 contour to synthesize emotional Mandarine speech from neutral. They argue that, speech variability is major concern in modifying prosody. The conversion must take place dependent of speaker than F0 contour itself [3]. It is a continuous approximation process. Later pitch targets are defined based on the equations below for syllable boundary [0,D].

$$T(t)= at+b \qquad (1)$$

$$y(t)= \beta \exp(-\lambda t) + a\, t + b \qquad (2)$$

$$0 \leq t \leq D, \lambda \geq 0$$

Here, *a* and *b* are slope and intercept of underlying pitch targets respectively. β measures the distance between pitch target and F0 contour for t=0. F0 is converted using Gaussian Mixture Models (GMM) and Classification and Regression Tree (CART). [4] Uses contours of pitch, energy and cepstral coefficients; that are continuously modelled over the duration of syllable. They represent the contour by representing it by its DCT (Discrete Cosine Transform) coefficients in its feature vector. It is having a benefit of unbound segment length of contour for mapping over the polynomial curve. Hirokazu Kameoka in his research paper defined a statistical model speech F0 contours. A version of Fujisaki model is formulated for discrete time stochastic process. Though expectation from statistical model is to frame out a powerful estimation of Fujisaki model, nature of speech pitch contours is represented in terms of probability distribution assumption. It was tested on artificially created data [5].

We can adapt the techniques used for speaker conversion with for the same prosody conversion. Both use pitch contours conversion. A good approach by taking DCT of the residual frames is narrated in [6]. By inverses filtering a speech signal the linear prediction residual is obtained from pitch synchronous frames. With the help of zero padding or trncating, dimension of DCT coefficient is modified in accordance with the desired factor. And then inverse DCT is obtained. Finally this period signal filtered forward to obtain modified speech signal. This was applied on affirmative sentences uttered in Kannada for concatenation based speech synthesis system. It is proven to be simple and elegant algorithm. We can also use emotion detection algorithms to model and modify the pitch contours. In [7], smoothing spline feature from speech signal.

[8] described a stochastic system which transforms pitch contours by taking into account multiple pitch parameters. Overlap-add (OLA) based system is used for the pitch conversion. They implemented a test environment on basis of pitch transplantation. Statistical algorithms are investigated in [9] for Korean language. It used Gaussian Normalization by combining it with declination-line modeling of pitch contours. The pitch contours are investigated at the different levels of accentual as well as intonation phrase. This method is proved to be the most accurate for modifying pitch contours even for large local pitch variations. So, we chose to go by Gaussian Normalization first followed by pitch contour statistical analysis and polynomial fitting.

III. GAUSSIAN NORMALIZATION

Mapping of neutral speech pitch contour F0 values to the desired is achieved by Gaussian Normalization technique. Expected value and standard deviation of the pitch contour vector of stressed words is calculated for both the desired interrogative and neutral speech in Marathi.

Algorithm: (A) Calculate and draw pitch contour for the desired and neutral speech. (B) Compute the mean and standard deviations for both. (C) Convert the pitch contour of desired stressed words of neutral speech pitch contour into the desired one by the formula given below on the basis frame-by-frame [MAIN].

$$x_2=(x_1-\mu_1/\sigma_1).\sigma_2+\mu_2 \qquad (3)$$

Here, μ and σ represent mean and standard deviation respectively.

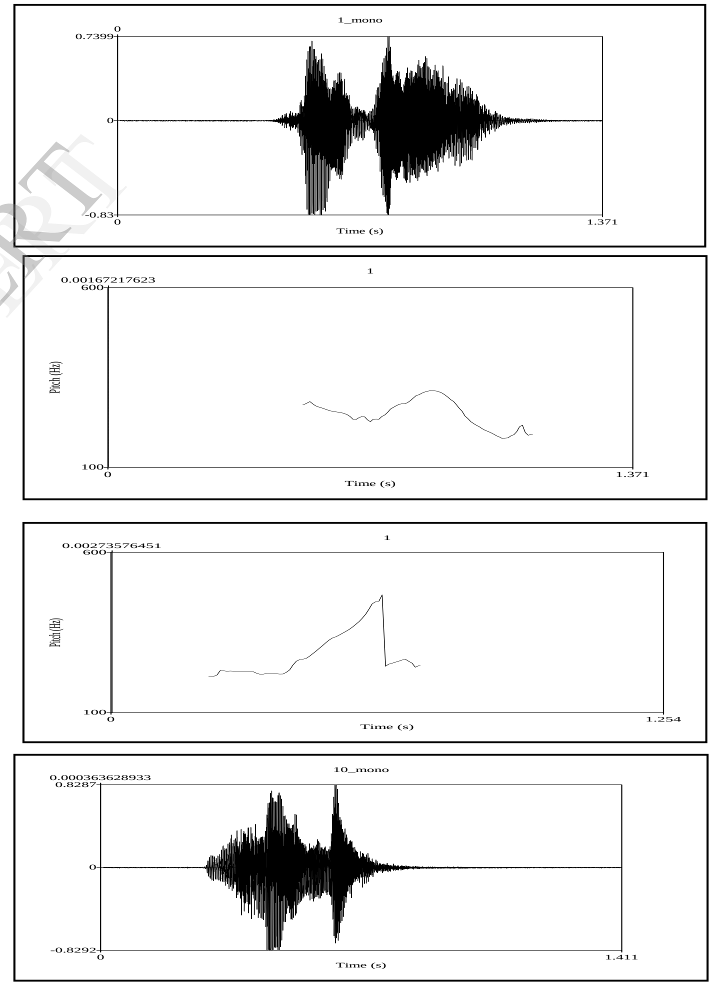

Fig.1. (a) Speech waveform of neutral Marathi word "havay" (b) neutral waveform's pitch contour (c) Pitch contour of the target emotion achieved after Gaussian normalization (d) retrieved sound waveform from the pitch contour in (c)

It can be easily perceived from the Fig. 1 that the pitch contour of neutral speech is mapped into desired interrogative word “*havay”* and is having different nature than that of desired. Subjective testing of the speech to before the native listeners of Marathi clearly gives the expected result. But, it is having a disadvantage of limitation of the word length. It is proven that it doesn’t work properly for log utterances. The only thing to be noted about it is the simplicity.

IV. POLYNOMIAL FITTING AND MAPPING FUNCTION

The above mentioned Gaussian normalization algorithm is generalized here. An important mapping function is developed form the neutral speech pitch contour and the desired one. Known pitch mappings are used for training followed by election of mapping function on the basis of best fit polynomial.

Algorithm: (A) Estimate the mean pitch for both the neutral and desired pitch contour for each voiced phone. (B) Construct a model of scatter-plot for each of the voiced phoneme. (C) Construct a data vector of each mean value calculated before. (D) Fit the scattered points into best fitting. (E) Use this mapping function for further mapping on frame by frame basis.

After testing for 1 to 10th order polynomial, it is observed that, the curve is best fitted with 4th order polynomial. There is performance variation. Sometimes algorithm gives non-linear mappings and many times its variation from linearity. If we plot it in the same manner, Gaussian normalization produces linear mapping response with control on intercept and slope of the line. By having a match of pitch values between desired prosody and neutral speech for same phones, a context level of sensation is introduced. In figure 2, the scatter plot of pitch values between target emotion and neutral speech. The sold line shows the best-fit 4thorder polynomial curve. Figure 3 shows the speech waveforms, neutral speech contour and modified contour achieved by using the mentioned algorithm. The algorithm is applied on the word “zala” in Marathi. It is observed that, while modifying the pitch contour, in Marathi, expressivity is not spread all over the sentence uttered. Instead of which, amongst all the words in a sentence, the last or first word is only needed to add the expressivity. So, for the neutral sentence in Marathi “pani piun zala”, expressivity is added into “zala” only, leaving the remaining word contour as it is. We have shown the only word to be stressed and whose pitch contour is to be modified.

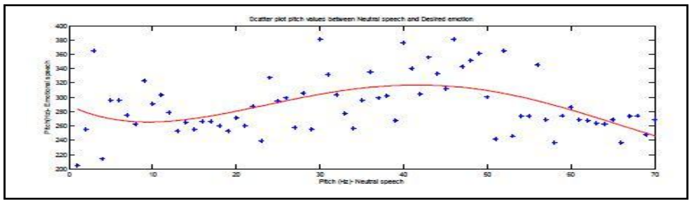

Fig. 2. A Scatterplot of neutral speech pitch values vs. desired emotion

.

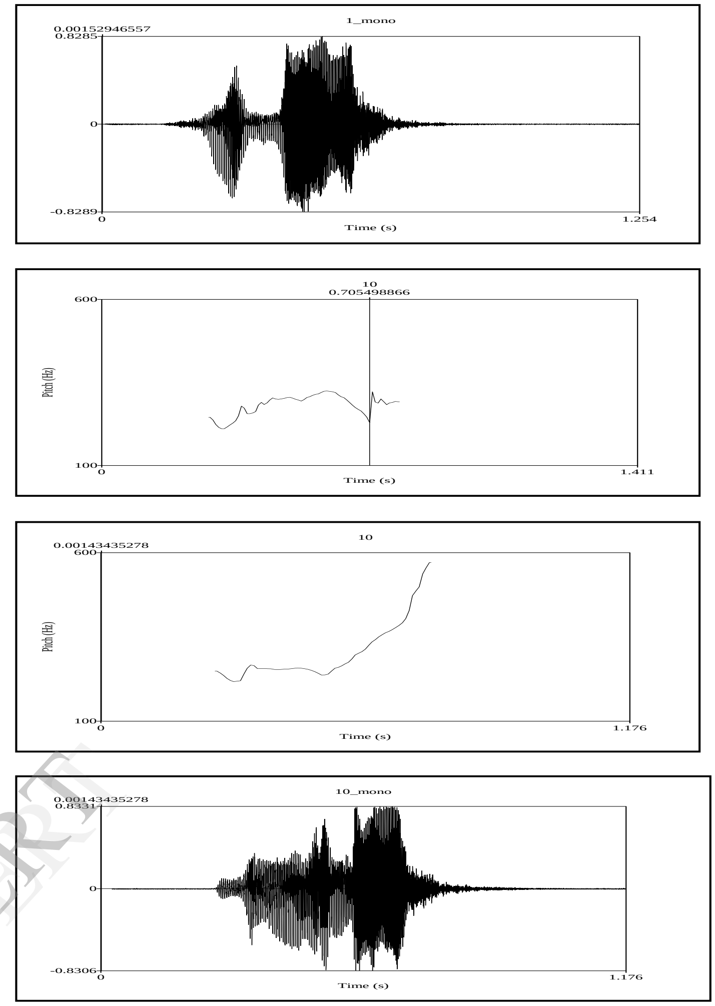

Fig 3. (a) Speech waveform of neutral Marathi word “zala” from the neutral Marathi sentence “pani piun zala” (b) Pitch contour if the word “zala” (c) Modified pitch contour according to the algorithm (d) Retrieved speech waveform.

This is somewhat language dependent as well as context dependent. When the speeches generated after modification was put for subjective testing before native listeners, they were satisfied for the modified prosody. Table 1 illustrates the pitch value analysis used for Marathi. The table helps in building model and ready data analyzed over large database.

TABLE I. Statistical analysis of pitch values in Hz.

MALE VOICE	
Minimum	103.27
Maximum	246.16
Mean	184.99
Standard Deviation	18%
FEMALE VOICE	
Minimum	204.50
Maximum	364.04
Mean	280.33
Standard Deviation	20%

V. RESULTS AND DISCUSSIONS

The experiment is done on 250 words, recorded in studio. They are sampled at 44100Hz for analysis; frame is selected of width 15msec. For the Gaussian normalization, the resulting couldn't possess much of the desired prosody. In most cases, the peaks in female voice are not satisfying while that of the male are satisfying. It doesn't give satisfactory performance for long utterances but gave well for short utterances. The scatterplot model performs better than the previous one. Listeners were much indulged than that of the Gaussian normalization by this technique.

As described in fig. 3, pitch contour of interrogative word is manipulated on the neutral Marathi word "zala" to get it converted into interrogative "zala". 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. 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 II 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.

VI. CONCLUSIONS AND FUTURE SCOPE

The Gaussian normalization mapping function is not suitable for the long utterances whose prosody is to be modified or changed. An abnormality is observed in the uttered word. So, this technique is suitable only for short duration words and utterances. But it is the simplest mapping function. It is easy to implement, computational complexity is less and so then time constraint is maintained. The scatterplot model for mapping is better than Gaussian normalization. It is the improved version of previous one. For some utterances it the perception of the speech was not clear but for most of them, it performed well. It is computationally somewhat complex and much of the time is required to analyse the data statistically.

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 is goaled.

TABLE II.

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

REFERENCES

[1] Jan P. H. Van Santen, "Prosodic modelling in Text-To-Speech synthesis", Lucent Technologies, Bell Labs, USA., pp. 2-4

[2] Yi Xu, Q. Emily Wang, "Pitch Targets and their realization: Evidence from Mandarine Chinese", Elsevier, Speech communication 33, 2001, pp. 319-337.

[3] Yongguo kang, Jianhua Tao, Bo Xu, "Applying pitch target model to convert F0 contour for expressive Mandarine speech synthesis", ICASSP 2006, vol. I, pp. 733-736.

[4] Marcel Kochkmann, Lukas Burget, "Contour modelling of prosodic and acoustic features for speaker recognition", SLT 2008, pp. 45-46.

[5] Hirokazu Kameoka, Jonathan Le Roux, Yasunori Ohishi, "A statistical model of speech F) contours", NTT Communication Science Laboratories, NTT Corporation, Japan, pp.1-3.

[6] R. Muralishankar, A. G. Ramakrishnan and P. Prathibha, "Modificationof Pitch using DCT in the Source Domain", Department of Electrical Engineering, Indian Institute of Science, Bangalore-560012, INDIA, pp.1-2.

[7] Frank Dellaert, Thomas Polzin and Alex Waibel, "Recognizing emotion in speech", School of Computer Science, Carnegie Mellon UniversityPittsburgh, Pennsylvania, pp.3-5.

[8] Tim Ceyysens, Werner Verhelst and Patrick Wambacq, "A strategy for pith conversion and its evaluation", Proc. 3rd IEEE Benelux Signal Processing Symposium (SPS-2002), Leuven, Belgium, March 2002, pp. S02-S04.

[9] Ki Young Lee, Yunxin Zhao, "Statistical Conversion Algorithms of Pitch Contours Based on Prosodic Phrases", Dept. of Information Communication Engineering, Kwandong University, S. Korea, Dept. of Computer Science, University of Missouri- Columbia, USA, pp. 1-3.