Prosody Generation in Neutral Speech for Storytelling Application Using Tilt Model

Manjare Chandraprabha A
Dept. of Electronics and Telecommunication
Rajarshi Shahu College of Engineering,
Pune, India.

S. D. Shirbahadurkar
Dr. D. Y. Patil college of Engineering
Ambi, Talegaon-Dabhade,
Pune, Maharastra, India

Anil Manjare
Toyo India Pvt. Ltd.
Mumbai, Maharastra, India

***Abstract*— This paper proposes Intonation Modeling for Prosody generation in Neutral speech for Marathi (language spoken in Maharashtra, India) story telling applications. Now a day's audio story telling devices are very eminent for children. Emotion puts precise meaning in the audio to understand content in speech. In this paper, we proposed tilt model for stressed words in Marathi for speech modification. Tilt model predicts modification in tone of neutral speech. GMM is used to identify stressed words for modification.**

***Index Terms*— Fundamental frequency, GMM, Statistical parametric Speech Synthesis, Tilt model.**

I.INTRODUCTION

Human storytellers use their voice in a variety of ways to capture their audience's attention. They mimic characters' voices, produce various 'sound effects' and use prosody to convey and invoke emotions, thus creating an engaging listening experience. In digital storytelling, stories are told by a computer. Ideally, a digital storytelling application should deliver a listening experience that is equally engaging as that provided by a human storyteller. To achieve this ideal we need a far more expressive, engaging speaking style than is provided by today's text-to-speech systems. However, we are currently working on improving these aspects of our system.

We focus on our improvements to the speech output of the Virtual Storyteller. To make the speech produced by the system more suitable for storytelling, we have focused on the creation of a general storytelling speech style. [2, 19] Many researchers has been worked on expressive speech synthesis which concentrates on basic emotions such as sadness, happiness, fear and anger. We need to concentrate on expressive speaking styles along with basic emotions. In this paper, we focus on a particular style that is storytelling speech.

[9] When compared to human speech, synthesized speech is distinguished by insufficient intelligibility, inappropriate prosody and inadequate expressiveness. These are serious drawbacks for conversational computer systems. Intelligibility is basic synthesis of intelligible phonemes is necessary for word recognition.[18] Prosody intonation (melody) and rhythm clarifies syntax and semantics, and aids in discourse flow control. Expressiveness, or affect, provides information about the speaker's mental state and intent beyond that revealed by word content. [20]Investigations into the perception of affect in human speech have isolated many of the acoustic and prosodic correlates of emotion. There has been research studied in which the hypothesis that the reproduction of these correlates in synthesized speech will result in perception of the original affect, and, that the manipulation of speech correlates of synthesized speech will produce changes in the perception of affect as predicted by the original psychoacoustic findings. To support or refute the hypothesis, it is necessary to consider the significant findings of psychoacoustic research, implement an affect synthesizer based on these findings and test its output with human subjects. [11, 12] Earlier, to attract audiences the human storytellers modified their voices. They use to create variety of voice and sound effects themselves to engage listeners. In an ideal case, story TTS should give a listening experience that is equally attractive as human story tellers. In digital storytelling stories are narrated by computer. Ideally digital storyteller should provide equal listening experience as of human storyteller. We want to create an application which will equal the voice of human storyteller by modifying the neutral voice parameters. To complete this target we need to concentrate on expressive style which is not available in the text to speech system these days.

II. RELATED WORK

Earlier speech synthesis of neutral speech is done with pitch contours [1] using HMM. Recently, the work was come up towards fully automatic extraction of Fujisaki Model Parameters [2].It has larger database of German. Fujisaki Model was used intended for production process of Fundamental frequency.

Some other papers [3]-[6], [8], [9] have used Formant synthesizers because of the high degree of control that they provide. These include the first ones, from 1989: Janet Cahn's Affect Editor [3], [4], and Iain Murray et al.'s HAMLET [5], [6]. Both have used DECtalk as a formant synthesis system, providing dedicated processing modules which adapt their input according to the acoustic properties of a number of emotions. In above both cases, the acoustic profile for each emotion category was derived from the literature and manually adapted. While the Affect Editor requires the input to be manually annotated, HAMLET processes its input entirely by rule.

IJSMS

Burkhardt, in his 2000 PhD [7]-[10], also used formant synthesis, despite the reduced naturalness, because of the high degree of flexibility and control over acoustic parameters that this technique provides. One approach to emotional speech synthesis with diphones, used in [11]-[16] papers. F0 and duration values are considered for each speech sound in a given utterance (usually an actor's portrayal of an emotion), and used for synthesizing the same utterance from diphones.

The result is a synthetic utterance with the same F0 and duration values as the actor's speech, but the voice quality determined by the diphones. This technique is suitable for modeling what humans do as closely as possible with the given parameter set.

III. DATABASE PREPARATION

The database is prepared for the purpose of analysis of neutral and emotional speech. The first step is to record the speech files for each word and sentence. The trained male and female are selected for the voice recording. Audacity software is used to record voice. [3] These words and sentences were recorded at a sample rate of 44100 Hz with a single (Mono) channel. The recordings were done in a closed, sound free room. For pitch analysis minimum and maximum pitch is taken as 75 Hz and 500 Hz. Speakers recorded the words and sentences, with each word and sentence being spoken in different prosodic emotions required for story telling application like fear, surprise, excitement, happy, sadness.

These words and sentences are emotionally rich and can be spoken in all five other emotions. Once the speaker is in emotionally charged mood, each speaker was asked to record the words and sentences with full emotion and stored in the computer. Normalization of speech means to adjust the volume so that the loudest peak is equal to the maximum signal that can be used in digital audio. Two methods were used for mitigating the background noise from the sound one is the manual spectral subtraction was done on each sound file by hand picking the noise regions and the second is the sound was subject to a noise gate to clean up the sound. As a last part of the optimization process, the files were trimmed so that silent regions at the beginning and end of each utterance were eliminated. The resulting database had an average duration for each sentence utterance.

IV. PROPOSED SYSTEM

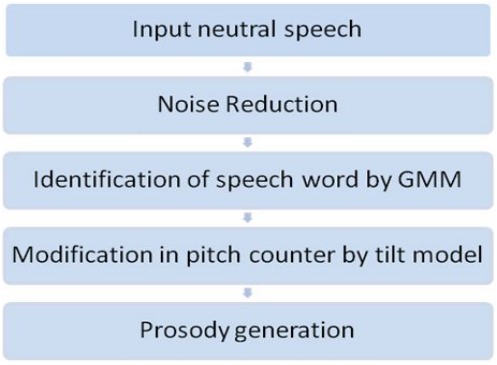


Fig. 1: Block diagram of proposed system to generate prosody

Our system introduces to define prosodic parameters to the tones for stressed words in Marathi story text. In this paper, Tilt model is proposed for speech synthesis. Three steps are involved in Tilt modeling. First step is to predict accent of stressed words. Second step is to generate to generate Tilt parameters. Third step is to generate F0 contour. The pitch contour represents the final pitch used to tone the particular word. The tone is determined by calculating "TILT".

Following parameters are used to calculate "TILT" [17] - [18]

a. Starting f0 value (Hz)
b. Duration
c. Amplitude of rise (A_{rise}, in Hz)
d. Amplitude of fall (A_{fall}, in Hz)
e. Starting point

The pitch contour shape, mathematically represented by its tilt, is a value computed directly from the f0 curve by the following formula:

$$tilt = \frac{|Arise| - |Afall|}{|Arise| + |Afall|} \quad (1)$$

$$tiltdur = \frac{|Drise| - |Dfall|}{|Drise| + |Dfall|} \quad (2)$$

Stressed words are identified with GMM model. GMM model is trained by the stressed words. Testing is done with the words recorded in different sound clips. GMM identifies stressed words. Tilt model modifies these words in the context of Amplitude and Duration.

V. RESULTS

Here Figs. 2 and 3 showing the speech signal for original neural and its emotional sentence. We required emotional speech signal of same sentence for comparative study purpose.

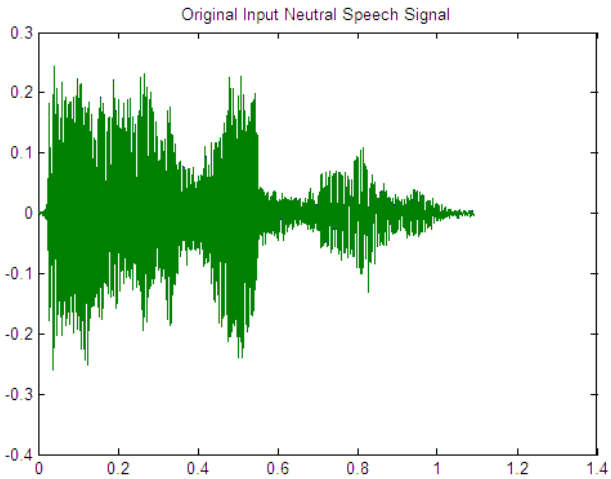


Fig. 2: Input Neutral Speech Signal

IJSMS

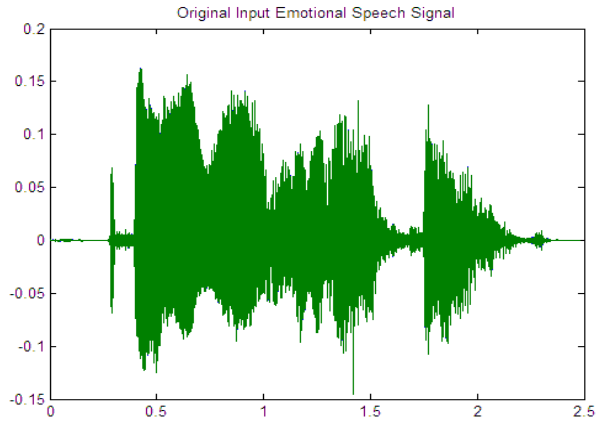


Fig. 3: Input Emotional speech signal for same statement

Table I: Ranking Used For Judging The Quality Of The Speech Signal
Rating Speech Quality

Rank	Speech Quality
1	No change in pitch
2	Poor
3	Moderate
4	Good
5	Excellent

Table II: Mean Opinion Scores for the Quality of Synthesized Speech

Emotion	Mean Opinion Score
Surprise	4.2
Fear	4.8
Excitement	4.1
Happy	4
Sadness	3.8

VI. COMPARATIVE ANALYSIS

In this section, we will present comparative analysis of neutral and modified speech in terms of two performance metrics such as Correlation (CORR) and Root Mean Square Error (RMSE) with below significance.

Correlation: This metrics represents the similarity between neutral speech and modified speech using cross correlation. Therefore, for efficient method it is required that less the correlation more the better method. Above results are showing the same for the proposed method.

RMSE: Root Mean Square Error is exactly opposite to Correlation metrics. This represents the difference between neutral speech and modified speech. Therefore, for efficient method it is required that RMSE should be more as compared to existing method.

We have tested proposed method on 10 input neutral speech signals and based on observations below Table III showing results of RMSE and CORR emotional and modified speech. In Table III, file name column shows ten tasted wave files.

Table III: Comparative Study

File Name	Correlation	RMSE
a.wav	0.8997	10.5173
b.wav	0.5189	10.5781
c.wav	0.7856	10.5873
d.wav	0.5547	10.5901
e.wav	0.6657	10.6244
f.wav	0.7093	10.6029
g.wav	0.8114	10.6232
h.wav	0.6858	10.6665
i.wav	0.9615	10.6211
j.wav	0.8367	10.5998

As per the significance stated above, our results are coming better showing comparison of emotional and modified speech.

VII. CONCLUSION AND FUTURE WORK

From the study of proposed method, results obtained for speech synthesis are better and improved as compared to existing methods. To overcome the limitations of existing methods such as F0 model, we have introduced new framework for speech synthesis called as F0 modification by Tilt model. We have prepared Marathi database for story telling sentences. GMM model is used to identify stressed words which express surprise, fear, excitement, happiness, and sadness. Amplitude and duration modification greatly affects on tone of neutral speech. For future work, we will suggest to work on speaker verification with speech synthesis under real time environment.

REFERENCES

[1] Gopala Krishna Anumanchipalli†‡, Lu´ıs C. Oliveira†, Alan W Black, "A Statistical Phrase/Accent Model for Intonation Modeling", INTERSPEECH 2011.

[2] J. Van Santen, A. Kain, E. Klabbers, and T. Mishra, "Synthesis of prosody using multi-level unit sequences," Speech Communication, vol. 46, pp. 365 – 375, 2005.

[3] Cahn, J. E., Generating Expression in Synthesized Speech, Master's Thesis, MIT, 1989.

[4] Taylor, P. , "Analysis and synthesis of intonation using the Tilt model," Journal of the Acoustical Society of America, vol. 107, no. 3, 2000, pp. 1697–1714.

[5] Cahn, J. E., The Generation of Affect in Synthesized Speech, Journal of the American Voice I/O Society, 8, July 1990, p. 1-19.

[6] Murray, I. R., Simulating emotion in synthetic speech, PhD Thesis, University of Dundee, UK, 1989.

[7] Murray, I. R., &Arnott, J. L., Implementation and testing of a system for producing emotion-by-rule in synthetic speech, Speech Communication, 16, p. 369-390.

[8] Montero, J. M., Gutiérrez-Arriola, J., Palazuelos, S., Enríquez, E., Aguilera, S., &Pardo, J. M., Emotional Speech Synthesis: From Speech Database to TTS, ICSLP 98, Vol. 3, p. 923-926.

[9] Burkhardt, F., Simulation emotionalerSprechweisemitSprachsyntheseverfahren [Simulation of emotional manner of speech using speech synthesis techniques], PhD Thesis, TU Berlin, 2000. http://www.kgw.tuberlin.de/~felixbur/publications/diss.ps.gz

[10] Burkhardt, F.,&Sendlmeier, W. F., Verification of Acoustical Correlates of Emotional Speech using Formant-Synthesis, ISCA

IJSMS

Workshop on Speech & Emotion, Northern Ireland 2000, p. 151-156.
[11] Mozziconacci, S. J. L., Speech Variability and Emotion: Production and Perception, PhD Thesis, Technical University Eindhoven, 1998.
[12] Mozziconacci, S. J. L., & Hermes, D. J., Role of intonation patterns in conveying emotion in speech, ICPhS 1999, p. 2001-2004.
[13] Vroomen, J., Collier, R., &Mozziconacci, S. J. L., Duration and Intonation in Emotional Speech,Eurospeech 93, Vol. 1, p. 577-580.
[14] Heuft, B., Portele, T., &Rauth, M. (1996), Emotions in Time Domain Synthesis, ICSLP 96.
[15] Rank, E., &Pirker, H., Generating Emotional Speech with a Concatenative Synthesizer, ICSLP 98, Vol. 3, p. 671-674.
[16] Montero, J. M., Gutiérrez-Arriola, J., Colás, J., Enríquez, E., &Pardo, J. M., Analysis and Modelling of Emotional Speech in Spanish, ICPhS 99, p. 957-960.
[17] Schröder, M., Can emotions be synthesized without controlling voice quality?, Phonus 4, Research Report of the Institute of Phonetics, University of the Saarland, p. 37-55. http://www.dfki.de/~schroed.
[18] Paul Taylor, University of Cambridge, Text-to-Speech synthesis.
[19] M. Schroeder, "Emotional speech synthesis: A review," Proceedings of Eurospeech 2001, 1, pp. 561-564, Aalborg, Denmark, 2001
[20] "Speech and Audio Processing" by Dr. Shaila D. Apte, Published by Wiley India Pvt. Ltd.

IJSMS