Modeling basic emotions for Tamil speech synthesis

A G Ramakrishnan and Lakshmi Chithambaran[#]

Department of Electrical Engineering, Indian Institute of Science, Bangalore
#Department of Electrical Engineering, M S Ramaiah Institute of Technology, Bangalore

Abstract: This paper explores the modeling of prosody parameters for improving naturalness of Tamil speech synthesis, by studying recorded utterances of speech with and without explicit emotions. To begin with, we look at interrogative and exclamatory Tamil sentences. Prosody parameters namely, pitch contour, energy and duration of each word in the sentences were observed, analyzed and generalized from the un-intonated and intonated, interrogative and exclamatory human speech. Differences in energy level were also analyzed between the two sets of utterances in three different frequency bands. Pitch is by modified in the LP residual domain using DCT. Energy is modified by multiplying the signal by the hypothesized factor and duration is modified as per the duration model by duplicating or removing integer number of pitch periods as necessary. The model was implemented on speech synthesized from Thirukkural TTS, developed by MILE LAB, and the results were found to be satisfactory.

Keywords: TTS, interrogative, exclamatory, pitch, energy, durations, LP, DCT, pitch synchronous, pitch modification, pitch contour, Gaussian modeling.

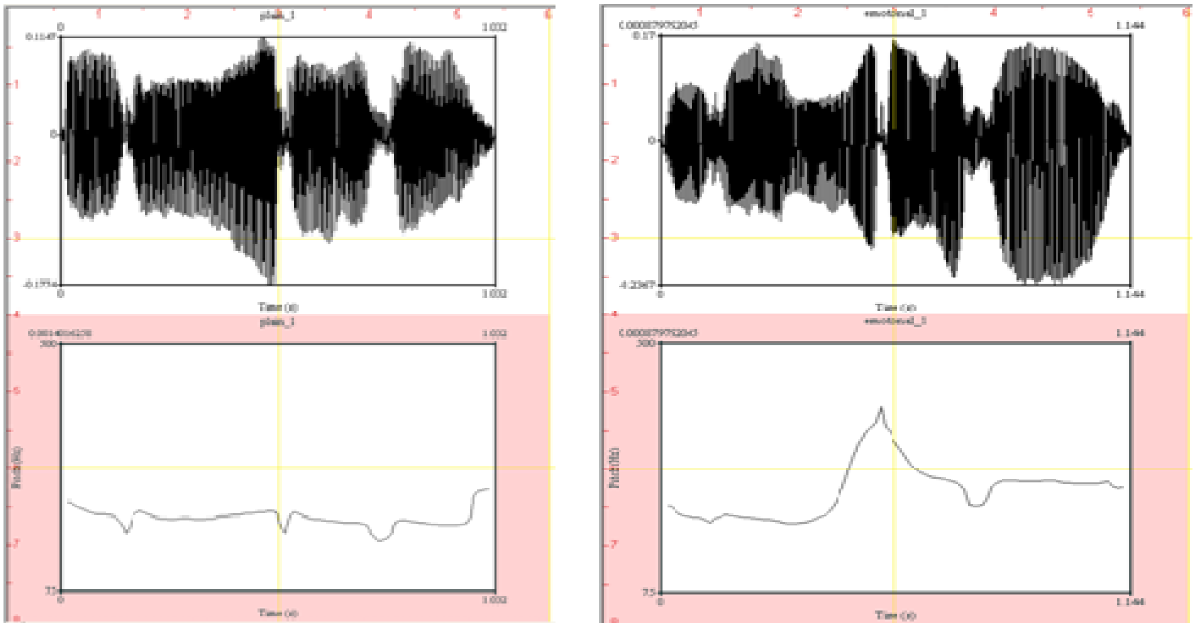
Introduction

It is known that a listener gets fatigued by listening to synthesized speech for a length of time [2]. This is fundamentally because text to speech conversion systems use normative speech from a native speaker to obtain their basic units for concatenation. Thus, it lacks variations in pitch, duration and energy level, which together are known as prosody [2, 3, and 4]. Further, a human being, even when she/he repeats the same sentence twice, the utterances are not identical, and there are always minute variations in stress levels and local amplitudes and duration. This lack of natural variations in speech attributes makes synthesized speech monotonous [3].

This triggers the need for modeling prosody in synthesized speech. Modeling prosody in languages like French, Spanish and English has a wide research literature. Modeling prosody in Indian languages like Tamil is not a trodden path. Here, interrogative and exclamatory intonations are considered. Prosody parameters are generalized for each of the above and appropriately modeled and implemented to speech synthesized by Thirukkural TTS, developed by MILE LAB.

Modeling interrogative intonation

Pitch contour: Figure 1 shows sample utterances of an interrogative sentence, both with and without intonation along with their pitch contours. It is noticed that the pitch contours of specific interrogation loaded words form a curve, with a rise and fall of pitch relative to the basal value. The mean ratio of maximum to basal pitch is 1.4. The question indicative word may fall in any position of the sentence: starting, somewhere in the middle or last. When the interrogative word (IW) is in the start or middle, it forms the rising part of the curve and the next word forms the falling part. Whereas, if the IW is the last word of the sentence, then it forms the falling part and the previous word forms the rising part of the curve. There are also interrogative sentences in Tamil, which do not have a specific IW; rather they end with the long vowel /aa/. In this case, the pitch significantly rises during the /aa/ sound.



Un-intonated interrogative speech and its pitch contour

Intonated interrogative speech and its pitch contour

Figure 1: Un-intonated and intonated interrogative speech and their respective pitch contours.

Energy: The energy and duration of the intonated interrogative utterance are higher by an average of 17.5% and 19.7%, respectively, than those of the un-intonated one.

Duration: The durations of the words that fall within the rise and fall region of the pitch contour also follow a distinct pattern. On the average, the duration of the first word in the region is 31.5% less than the un-intonated word, whereas that of the second word is 32% more.

Pitch modification is carried out pitch synchronously by interpolation of LPC residues of each pitch period in the DCT domain. Energy modification is accomplished by multiplying the signal by the hypothesized factor. Durational is modified by duplicating or removing integral number of periods. Tables 1 and 2 compare the mean opinion scores of five natives on un-intonated human speech and TTS speech, respectively before and after the prosodic modifications.

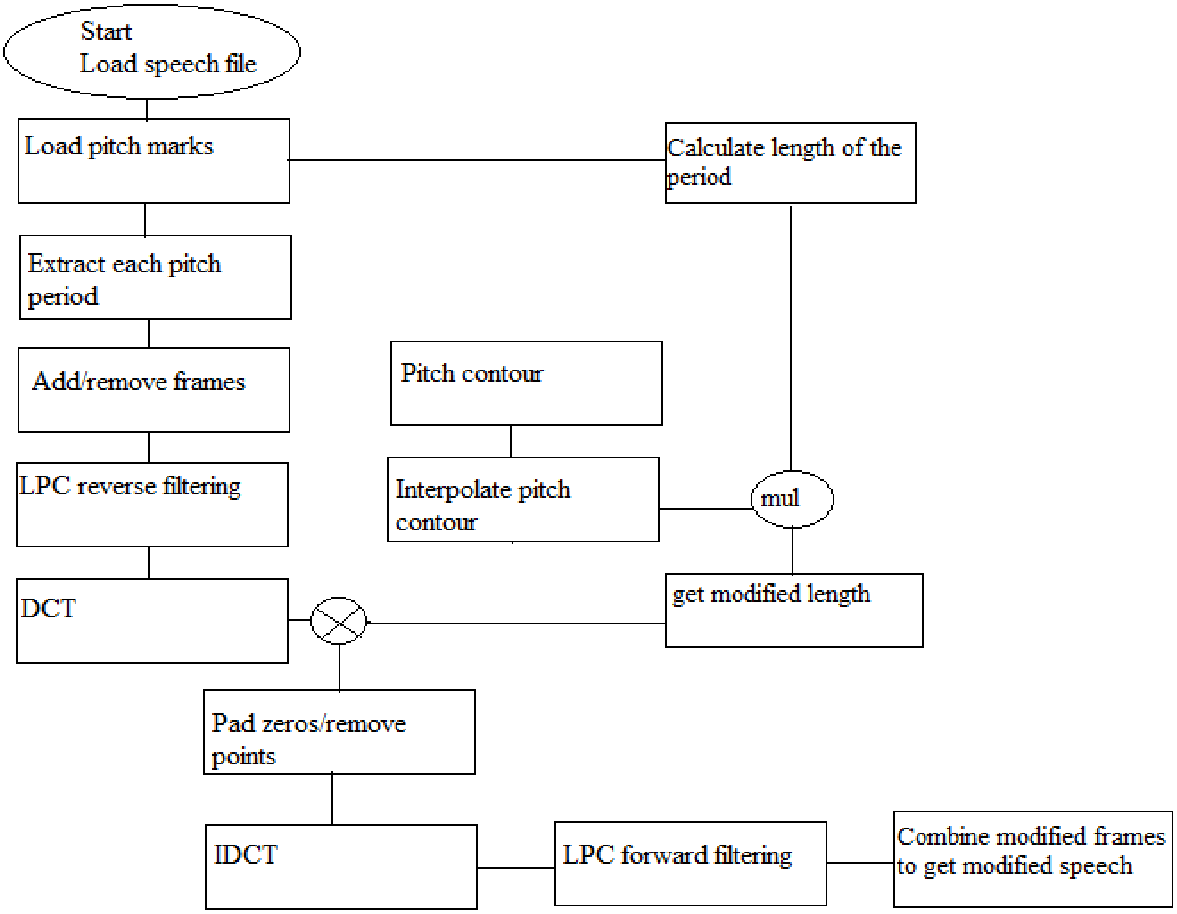


Figure 2: Block diagram implemented for pitch modification as per hypothesis.

Table 1: Comparison of MOS (on a scale of 5) of un-intonated human speech after modification.

Type of speech	MOS
Sentence recorded without intonation	2
Un-intonated recording modified by our algorithm	3.8
Sentence recorded from a human with intonation	5

Table 2: Comparison of MOS of TTS speech after prosody modification.

Type of speech	MOS
TTS output	**1**
TTS output modified by new algorithm	**2.8**
Sentence recorded with intonation	**5**

Contour Analysis: It is noticed that noise in the modified speech is due to the random pitch contour present in the TTS speech. To nullify this effect, pitch contour of the TTS speech is analyzed and if the contour by itself satisfies the hypothesized values within an error of ± 5%, then the TTS is not processed at all; else, the TTS speech is processed as discussed.

Table 3: Comparison of MOS of TTS speeches as per contour analysis

Type of speech	MOS
TTS output	**1**
TTS output modified by our new algorithm	**3.5**
Sentence uttered with intonation	**5**

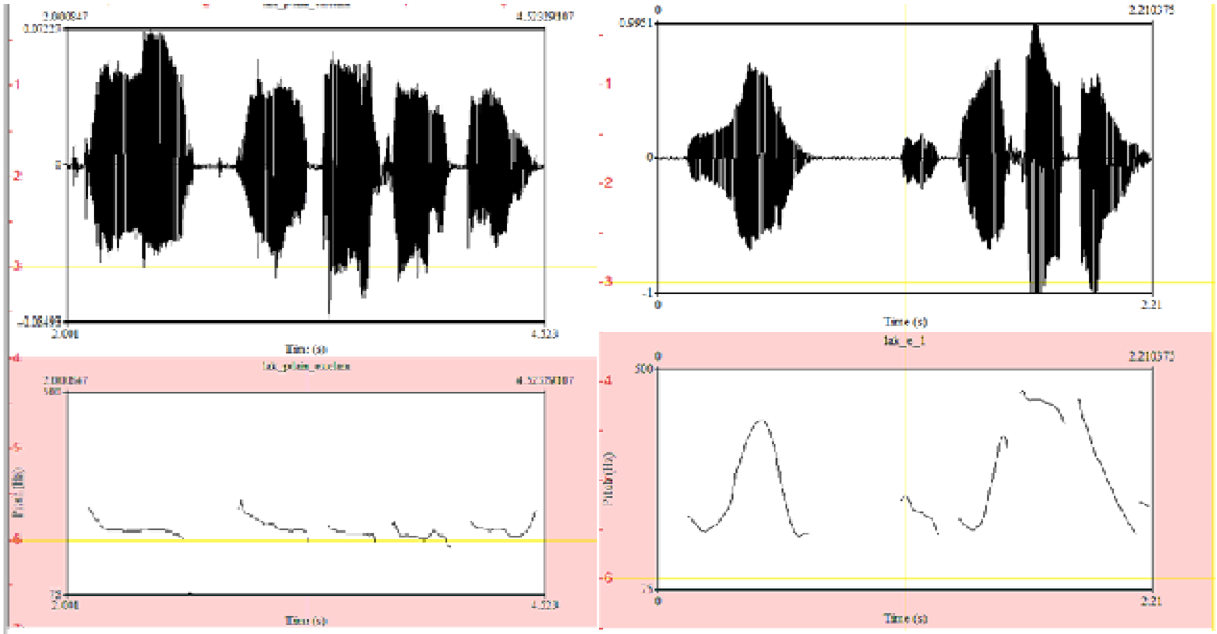
Energy Distributions: The fractions of energy in the bands, 0-500 Hz (low frequency band), 500-2000 Hz (mid band) and 2000-8000 Hz (high band) of un-intonated and intonated speech are analyzed. For the first word that falls within the curve, the fractions of energy in low and mid bands of intonated speech are higher on an average by a factor of 3.2 than their un- intonated counterparts. For the second word, the energy in mid frequency band of intonated speech is higher by a factor of 5.43 than their un-intonated counterpart. There is no significant change in the high frequency band for either word. Therefore, by scaling the energies in the respective frequency bands only, we obtain better intonation and clarity of words in the modified speech.

Table 4: Comparison of MOS of TTS speeches as per energy distribution analysis.

Type of speech	MOS (emotion)	MOS (clarity)
TTS output only with pitch modification (TOPM)	**3**	**3.3**
TOPM and increase of energy of the signal	**3.6**	**3.6**
TOPM and scaling of energy only in specific bands	**3.5**	**3.8**

Modeling exclamatory intonation:

It is noticed that the pitch contour of the last two words in the sentence takes a rise and fall pattern. The mean ratio of maximum to basal pitch is 1.8.



Un-intonated exclamatory sentnces and its pitch contour.

Intonated exclamatory sentnces and its pitch contour.

Figure 3: Un-intonated and intonated exclamatory speech and their respective pitch contours.

The energy and duration of the intonated exclamatory sentence are higher by an average of 19.3% and 16.4% than those of un-intonated one, respectively. The pitch, energy and duration are modified in the same way as the interrogative case. Table 5 shows the evaluation of the results of the modifications.

Table 5: Comparison of MOS of un-intonated exclamatory speeches.

Type of speech	MOS
Sentence recorded without intonation	1
Un-intonated recording modified by our algorithm	4
Sentence recorded with intonation	5

Energy in specific bands: As discussed for the interrogative case, energy in different frequency bands are analyzed for exclamatory intonation as well. For the first and second words within the pitch curve, the energies in mid band of intonated speech are 6.25 and 5.4 times those of un-intonated speech, respectively. Change in high frequency band is not significant for both words. Improvement in the quality of processed speech is evaluated and listed in Table 6.

Table 6: Comparison of MOS of TTS speeches as per energy distribution analysis.

Type of speech	MOS (emotion)	MOS (clarity)
TTS output only with pitch modification (TOPM)	**3.75**	**3.5**
TOPM and increase of energy of the signal	**4.25**	**3.75**
TOPM and scaling of energy in specific bands	**4.2**	**4**

Results and Discussion

Prosody of interrogative and exclamatory sentences have been modeled. Modification of the pith, energy and duration of the TTS generated speech gives interrogative or exclamatory intonation, as evaluated by the native listeners. Energy modification in certain frequency bands gives better quality and intended intonation closer to expected natural intonation.

Acknowledgment: The authors thank Mr. Shiva Kumar HR for creating the MILE TTS web demo, used for the experiments reported in this paper.

References

1. Fabio Tamburini, “Prosodic prominence detection in speech,” Proc. VII ISSPA 2003.
2. G.L.Jayavardhana Rama, A.G.Ramakrishnan, M.Vijay Venkatesh, and R.Muralishankar, “Thirukkural - a text-to-speech synthesis system,” Proc. Tamil Internet 2001, Kuala Lumpur, August 26-28, 2001, pp. 92-97.
3. R. Muralishankar and A. G. Ramakrishnan, “Human touch to the Tamil Synthesizer,” Proc. Tamil Internet 2001, Kuala Lumpur, August 26-28, 2001, pp. 103-109.
4. R Muralishankar, A.G.Ramakrishnan and P Prathibha, “Modification of pitch using DCT in the source domain,” Speech Communication, 2004, Vol. 42/2, pp. 143-154.
5. R. Muralishankar and A. G. Ramakrishnan, “Synthesis of speech with emotions,” Proc. Intern. Conf. Commn., Computers and Devices, Kharagpur, Dec. 14-16, 2000, pp. 767-770.
6. G. L. Jayavardhana Rama, A. G. Ramakrishnan, R. Muralishankar and P. Prathibha, “A Complete Text-to-Speech Synthesis System in Tamil,” Proc. IEEE 2002 Workshop Speech Synthesis, Santa Monica, CA USA, Sep. 11-13, 2002, pp. 191-194.
7. A. G. Ramakrishnan, Lakshmish N Kaushik and LaxmiNarayana M, “Natural Language Processing for Tamil TTS,” Proc. 3rd Language and Technology Conference, Poznan, Poland, Oct 5-7, 2007, pp. 192-196.
8. L. R Rabiner and R. W Schafer “Digital Processing of Speech Signals”, PHI, 1978.
9. Web demo of MILE TTS: http://mile.ee.iisc.ernet.in/tts