Human Touch to Tamil Speech Synthesizer

R Muralishankar and A G Ramakrishnan
Department of Electrical Engineering, Indian Institute of Science, Bangalore-12

INTRODUCTION

The continuous increase in synthetic speech intelligibility has shifted the focus of attention to naturalness. One of the important aspects of providing naturalness to the synthesized speech involves a great deal of study of the duration of each segment in naturally uttered words. This information can be put into the lookup table and it is very useful in deciding the duration of each segment before concatenation. Apart from the duration, mimicking the diversity of natural voice is the aim of many current speech investigations. The emotional expression in the synthesized speech is achieved by modifying the prosody (intensity, duration and pitch) appropriately.

NATURAL-SOUINDING SYNTHETIC SPEECH

As far as the listener is concerned, natural-sounding synthetic speech is, by definition, indistinguishable from real speech. This does not mean that the synthetic speech is exactly the same as real speech. Current theories of language and speech are not adequate to enable us to replicate speech production. The goal therefore is to produce a simulation of the human output, which is perceptually accurate, by employing a system, which is as good a simulation as we can manage of the human processes, which derive that output.

Human speech is characterised by (1) segmental and (2) suprasegmental features, which collectively contribute to the naturalness of speech. A segment refers to some small chosen unit of speech (example. phonemes and syllables). Segmental features refer to those, which decide the phonetic quality of the segment. Suprasegmental or prosodic features have their domain extended over more than one segment i.e., syllables, morphemes, phrases, sentences, etc. The suprasegmental features, namely, the rhythm (duration), stress (intensity) and intonation (pitch) are influenced by the factors such as phonetic, syntactic context, semantics & emotional state of the speaker. The terms in parentheses are the acoustic parameters in which the corresponding features are manifested [1]. Suprasegmental features are the overlaid functions of the corresponding segmental features [2].

USER REACTION TO POOR SYNTHETIC OUTPUT

Current synthesis systems often produce voice output, which sounds monotonous, unnatural and is tiring to listen to. The speech produced cannot be listened to easily over periods of time even as short as a paragraph span. In an interactive dialogue situation users become irritated with system, and in other situations such as where the system is giving instructions, the user can become bored or uninterested.

Good speech output is important for dialogue systems because user awareness is heightened in dialogue mode: the listener's attention is focused on the speech output, since the task of decoding speech requires concentration. In addition to plain message, all of the information about the thoughts, ideas and feelings that are being communicated is encoded in the speech waveform, and the range of variability in natural speech is narrow. Common errors in current synthesis systems are poor quality, limited bandwidth, inadequate segment conjoining, monotonous and inappropriate intonation and poor stress assignment.

LACK OF NATURALNESS IN SYNTHETIC OUTPUT

There are a number of factors, which contribute to the lack of naturalness in the speech output from speech synthesis systems.

(a) Intonation and rhythm

Errors of intonation and rhythm lead to monotonous or incorrect output, or can contribute to a misunderstanding of the meaning of what is being said. Intonation errors arise from inadequately modelling intonation generation, incorrect assignment of prosodic markers at a higher linguistic level, and incorrect interpretation of these markers at lower levels. Errors of rhythm arise from failure to model adequately the way in which segmental durations vary during an utterance by failing to set an appropriate range of acceptable variation.

(b) Variability along the prosodic parameters

Another source of error involving the prosodic parameters of duration and the fundamental frequency is the failure to take into account the fact that in human speech, these parameters vary for specific effects. For example, in order to focus the listener's attention on a specific word or phrase, one slows down the overall rhythm. A speaker may pitch the overall fundamental frequency a little lower to indicate that the current piece of information is confidential between speaker and listener, and not intended for anyone else (even if no one else is currently present); this is often accompanied by an overall drop in acoustic amplitude.

(c) Incorrect segmental rendering

Errors generated in the phonological processing in a synthesis system can lead to an incorrect choice of segments for rendering part of a particular word. There are, for example, occasions in human speech where vowel reduction under stress conditions or in slow speech is not correct. In fast speech, on the other hand, there may be occasions where greater vowel reduction is called for in unstressed syllables, or even total deletion of these syllables.

Fig. 1 shows schematically the elements of prosodic generation for a natural speech synthesis system. Prosody depends not only on the linguistic content of the signal. Different people generate different prosody depending on their mood. The speaking style of the voice can impart an overall tone to a communication. Narrowed pitch range indicates boredom, depression, or controlled anger. Another example of a global effect is a very fast speaking rate that might signal excitement. Examples of local effects are a notable excursion of pitch,

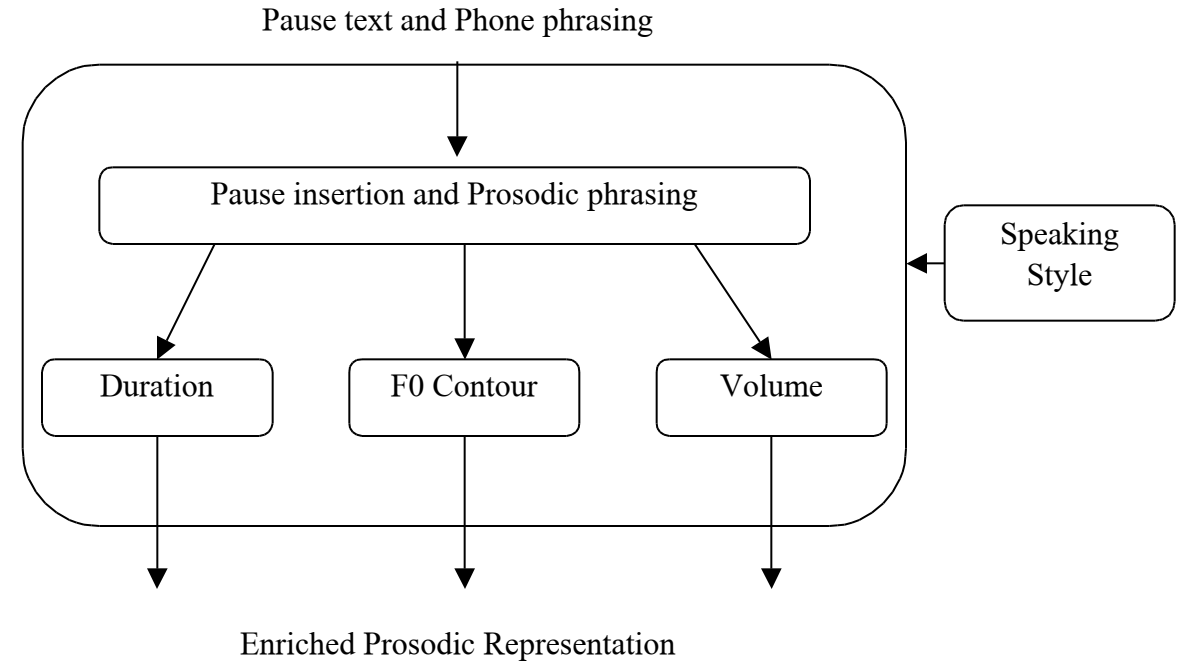

Fig. 1 Block diagram of a prosody generation system.

higher or lower than surrounding syllables, for a syllable in a word chosen for special emphasis. Another example would be the typical short, extreme rise in pitch on the last syllable of a yes-no question. Figs. 2(a) and b (top panel) show the waveforms for the utterance "Avan Nejamma Varaan" spoken with two different intonations. The bottom

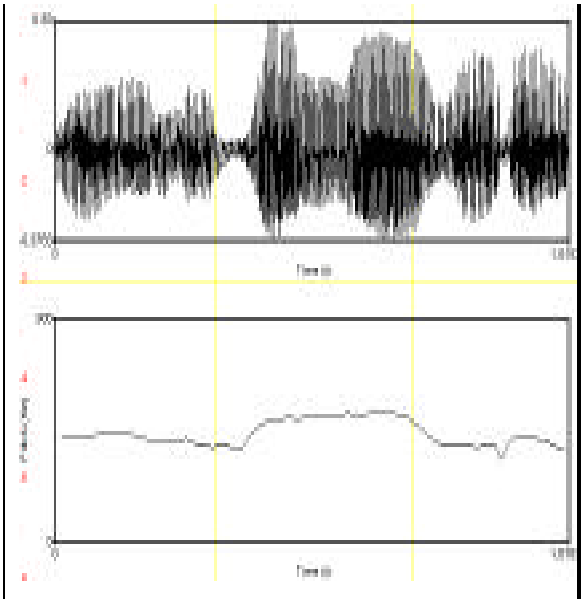

Fig. 2(a) Affirmative utterance of a sentence and its pitch contour

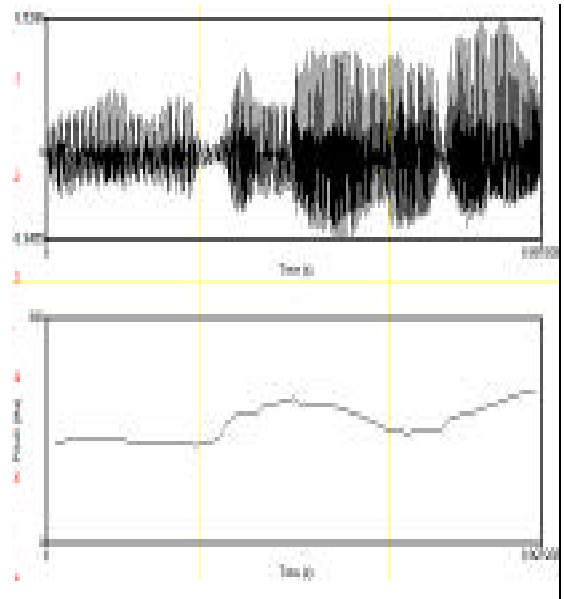

Fig. 2(b) Interrogative utterance of the same sentence and its pitch contour

panel of the figures show the pitch contour obtained by analysing the above waveforms [3]. In Fig. 2(a), the absolute pitch decreases towards the end, which is typical of affirmative sentences. In Fig. 2(b), the pitch increases towards the latter part of the sentence, indicating that it is an interrogative sentence.

Temporary emotional conditions such as amusement, anger, contempt, grief, sympathy and suspicion too have an effect on prosody. Just as a film director explains the emotional context of a scene to her actors to motivate their most convincing performance, a speech synthesis system needs to provide information on the simulated speaker's state of mind. These are relatively unstable properties. That is, one could imagine a speaker with any combination of social/dialect/gender/age characteristics being in anyone of a number of emotional states that have been found to have prosodic correlates, such as anger, grief and happiness. An additional complication in expressing emotion is that the phonetic correlates appear not to be limited to the major prosodic variables (F0, duration, energy) alone. Besides these, effects in the voice such as jitter (inter-pitch-period microvariation), or the mode of excitation may be important [4]. In a formant synthesizer supported by extremely sophisticated controls [5], and with sufficient data for automatic learning, such voice effects might be simulated. In a typical time-domain synthesizer, the lower level phonetic details are not directly accessible, and only F0, duration, and energy are available.

Pitch and duration are not entirely independent, and many of the higher-order semantic factors that determine pitch contours may also influence durational effects. The relation between duration and pitch events is a complex and subtle area, in which only initial exploration has been done [6]. Nonetheless, most systems often treat duration and pitch independently because of practical considerations [7]. Numerous factors, including semantics and pragmatic conditions, might ultimately influence phoneme durations. Some typically neglected factors include: the issue of speech rate relative to speaker intent, mood, and emotion, the lack of a consistent and coherent practical definition of the phone such that boundaries can be clearly located for measurement. This clearly shows that the durational information of the target word in a sentence should be known clearly before the concatenation of the basic units like V, CV, VCV, VCCV and VCCCV as considered in Thirukkural [8]. Fig. 3 shows the variation in the duration of the word, 'sendraan' when it is uttered in isolation, as the final word of a sentence, and when it is in the middle of a sentence. This clearly shows the need for modification of duration to obtain naturalness.

METHOD

Presently, we are working on incorporating techniques like synthesis of speech emotion [9], pitch modification using DCT (Discrete cosine transform) [10] and durational rules into the Tamil speech synthesizer [8]. Here, all these techniques are applied pitch synchronously. For synthesizing speech with emotions, we have the database of pitch contour and spectral information obtained by linear prediction (LP) analysis, for several natural emotions. This information has been used to modify the pitch contour of inverse filtered, neutral segment and after forward filtering it with the LP parameters to synthesize the given emotion.

Utterance	Duration of "சென்றான்"
சென்றான்.	820 mSec
அவன் சென்றான்.	687 mSec
அவனும் சென்றான், அவளும் சென்றாள்.	631 mSec
அவன் சென்றான் எனக் கேள்விப்பட்டேன்.	489 mSec

Fig. 3. Variation in the duration of a word in different circumstances.

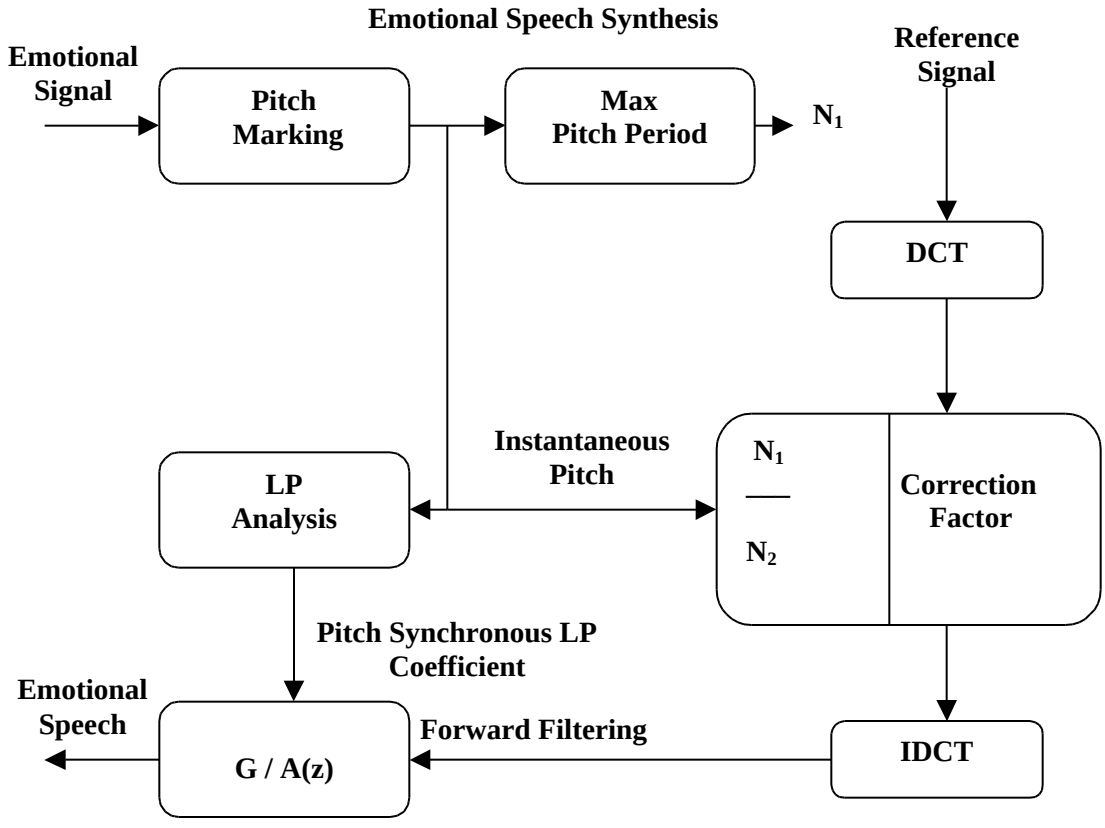

Fig. 4. Block Schematic of Emotional Speech Synthesis

Fig. 4 shows the block diagram of the scheme for emotional speech synthesis. The above technique has been used in the synthesizer to speak out interrogative and exclamatory sentences. With the pitch-marking algorithm, we can add or subtract the number of cycles

required to suit the duration of the segment for concatenation. Pitch modification-using DCT [10] has been used here to raise or lower the pitch contour of the segments before concatenation, so that the distortion due to the pitch mismatch between the segments is minimised. The block diagram of the scheme for pitch modification is given in Fig. 5. This improves the intelligibility as well as naturality.

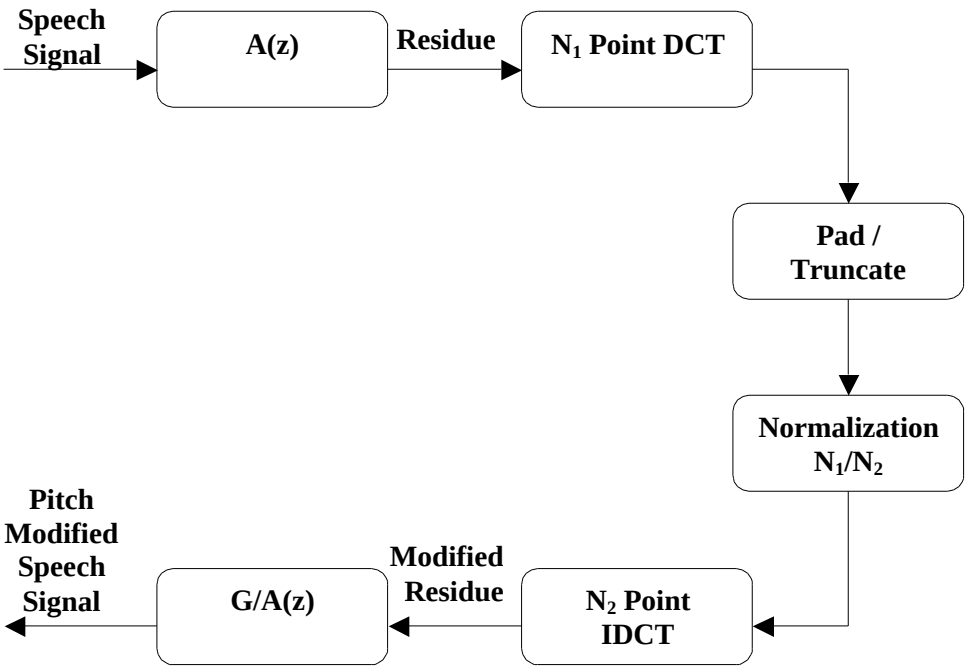

Fig. 5. Block Schematic of Pitch Modification System

RESULTS

Figure 6a displays the spectrogram of the original utterance shown in Fig. 2. The spectrogram of the same utterance after pitch modification by a factor of 1.2 is illustrated in

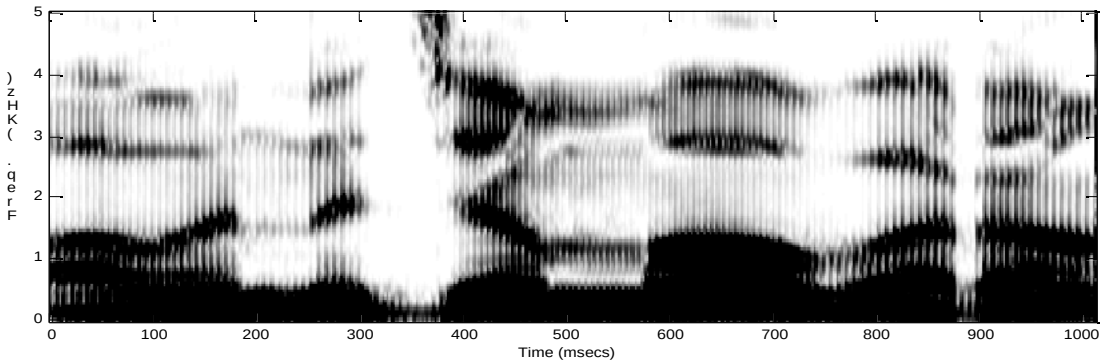

Fig. 6b. As is evident from the figures, the shape of the formant contour is unchanged.

Fig. 6(a) Spectrogram of the utterance shown in Fig. 2

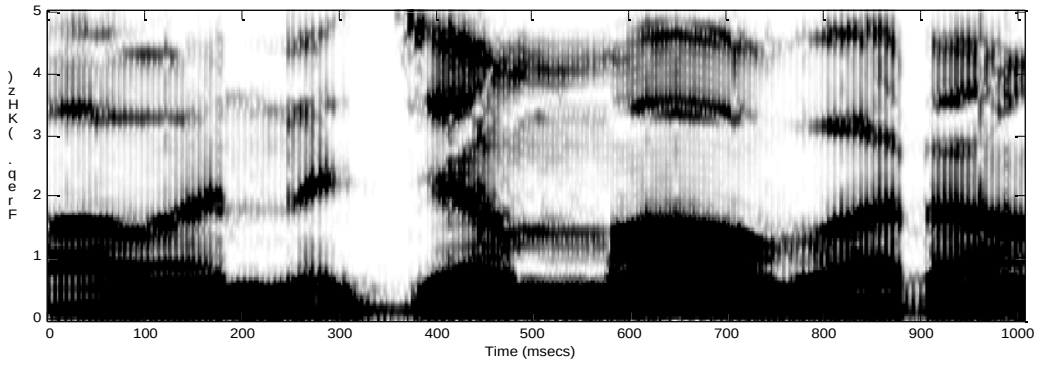

Fig. 6(b) Spectrogram of the utterance after pitch modification by a factor of 1.2.

REFERENCES

[1] I Lehiste, Suprasegmentals, The MIT Press, Cambridge, 1970.

[2] S R Rajesh Kumar, "Significance of durational knowledge for a text-to-speech system in an Indian language," M.S. Thesis, I. I. T., Madras, March 1990.

[3] R. Muralishankar and A G Ramakrishnan, "Robust Pitch detection using DCT based Spectral Autocorrelation", Proc. Intern. Conf. on Multimedia Processing, Chennai, Aug. 13-15, 2000, pp. 129-132.

[4] Klasmeyer, G. and W.F.Sendlmeier, "The Classification of different phonation types in emotional and neutral speech," Forensic Linguistics, 1997, 4(1), pp.104-125.

[5] K. N. Stevens, "Control Parameters for Synthesis by Rule," Proc. of the ESCA Tutorial Day on Speech Synthesis, 1990, pp. 27-37.

[6] J. van Santen, and J. Hirschberg, "Segmental Effects of Timing and Height of Pitch Contours," Proc. Int. Conf. on Spoken Language Processing, 1994. pp. 719-722.

[7] J. van Santen, "Assignment of Segmental Duration in Text-to-Speech Synthesis," Computer Speech and Language, 1994, 8, pp. 95-128.

[8] G L Jayavardhana Rama et al. "Thirukkural : A Speech synthesis system in Tamil", accepted for presentation, Tamilnet2001, Kualalumpur, Malaysia, Aug. 26-28, 2001.

[9] R. Muralishankar and A G Ramakrishnan, "Synthesis of Speech with Emotion", Proc. Intern. Conf. on Commn., Computers and Devices, Vol. II, Kharagpur, Dec. 14-16, 2000, pp. 767-770.

[10] R. Muralishankar et al. "DCT based Pitch Modification", accepted for oral presentation in Sixth Biennial Conference on Signal Processing and Communications, IISc, Bangalore, July 15-18, 2001.