A COMPLETE TEXT-TO-SPEECH SYNTHESIS SYSTEM IN TAMIL

G. L. Jayavardhana Rama, A. G. Ramakrishnan, R. Muralishankar and P. Prathibha

Department of Electrical Engineering, Indian Institute of Science, Bangalore - 560012, INDIA.

ABSTRACT

We report the design and development of *Thirukkural*, the first text-to-speech converter in Tamil. Syllables of different lengths have been selected as units since Tamil is a syllabic language. Automatic segmentation algorithm [8] has been devised for segmenting syllables into consonant and vowel. The units are pitch marked using Discrete Cosine Transform - Spectral Auto-correlation Function (DCT-SAF) [6]. Prosodic information is captured in tables based on extensive observation of spoken Tamil. During synthesis, DCT based pitch modification [3][7][11] is applied for both waveform interpolation and modifying pitch contour for different sentence modalities. Thirukkural is designed in VC++ and runs on windows 95/98/NT. Perceptual evaluation by natives show that the synthesized speech is intelligible and fairly natural.

1. INTRODUCTION

Over 65 million people worldwide speak Tamil, the official language of the southern state of Tamil Nadu, and also of Singapore, Sri Lanka and Mauritius. In addition to the above countries, it is spoken in Bahrain, Malaysia, Qatar, Thailand, United Arab Emirates and United Kingdom. It will be a boon to Tamilian, if the user interface with the computer is in Tamil, that too if it is in the form of speech. Giving such a natural language interface to the user is one of the greatest applications of a text to speech converter. The synthesizer can be used as an automatic text reader for the blind [4]. In telecommunication it can be used for reading e-mails, web pages and also as talking newspaper. It can be used as a multimedia teaching aid. In this paper, we describe *THIRUKKURAL*, a Text-to-Speech (TTS) converter for Tamil [2].

2. INTRODUCTION TO TAMIL LANGUAGE

Tamil alphabet has descended from the Brahmi script of ancient India. It is a syllabic language which contains 12 vowels and 18 consonants in its script. The language has certain well defined rules which introduce seven other phones depending on the presence of consonants with respect to the vowels or the other consonants. Hence there are 39 phones in the language. The rules which are made use of for synthesis are as follows:

- When a hard consonant is followed and preceded by a vowel, automatically the hard consonant acquires the softer version
- When a hard consonant is preceded by its nasal, hard consonant gets converted to soft consonant
- /ch/ is pronounced as /s/ when it comes in the beginning of a sentence

3. OVERVIEW OF THIRUKKURAL

Figure 1 shows the flow chart of the text to speech conversion in Thirukkural. Concatenation with waveform modification has more flexibility in selecting the speech segments to concatenate because the waveforms can be modified to allow for a better prosody match [1]. This means that the number of sentences with mediocre quality is lower than the case where no prosody modification is allowed. On the other hand, replacing natural with synthetic prosody can hurt the overall quality. In addition, the prosody modification process also degrades the overall quality. Analysing the advantages and disadvantages of concatenation with waveform modification, we proceed to propose a method to improve the quality of the synthesized speech. Our system performs waveform modification to achieve naturalness and to convey the sentence modality (affirmative, interrogative or exclamative). The system can be broadly classified into two phases namely the offline phase and online phase. Offline phase includes pre-processing, segmentation and pitch marking. Online phase includes text analysis and synthesis.

3.1. OFFLINE PROCESS

The offline processes of the system include (1) Choosing the basic units (2) Building the database (3) Detailed study of prosody in natural speech (4) Consonant - vowel segmentation (5) Pitch marking.

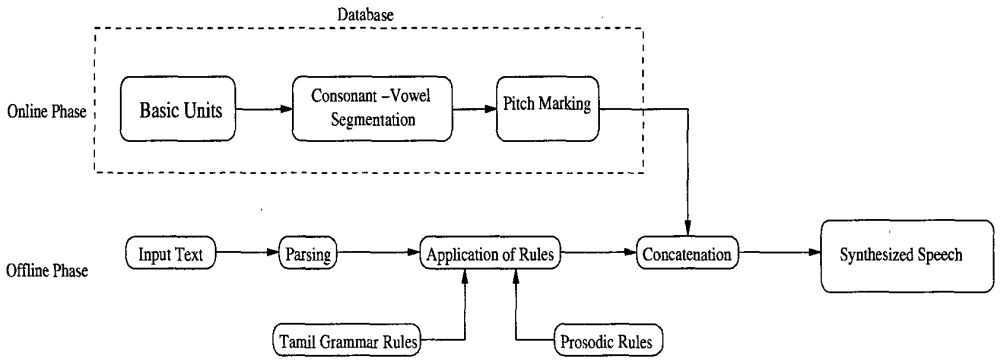

Fig. 1. Flow chart of text to speech conversion in Thirukkural

3.1.1. Deciding the basic units

The issues in choosing the basic units [5] for synthesis are:

- The units should lead to low concatenation distortion.
- The units should lead to low prosodic distortion.
- The units should be of a general nature, if unrestricted text-to-speech is required.

Considering the above issues, syllables have been used as basic units. This may contain phonemes, diphones or triphones. The different instances of the unit are V, CV, VC, VCV, VCCV and VCCCV, where V stands for a vowel and C stands for a consonant. Since duration rules are applied during concatenation, the database contains only long vowels. The total number of units is around 3000.

3.1.2. Building the database

The database was collected from a native Tamil speaker over a span of one month. Recording took place in a noise free room using Shure SM 58 microphone, whose frequency response is 50-15,000 Hz and SNR is 30 dB under lab conditions. Spoken units were recorded at a sampling rate of 8 KHz.

3.1.3. Observation of prosody in natural speech

Prosody is a complex weave of physical, phonetic effects that is being employed for expression. Prosody consists of systematic perception and recovery of the speaker's intentions based on pauses, pitch, duration and loudness. Pauses are used to indicate phrases and to indicate the end of sentences or paragraphs. It has been observed that the silence in speech increases as we go from comma to end of sentence to end of paragraph. Pitch is the most expressive part of a speech signal. We try to express our emotion through pitch variation. Speech with constant pitch sounds very unnatural. In Tamil, sudden variation in pitch does not occur in a vowel as it happens in Japanese. Duration is the second important factor that affects the naturalness of the synthesized speech. Same vowel appearing in different positions in a word or a sentence has different durations. For example, consider the sentence "/naan aaru manikku varalaamaa/". The duration of vowel /aa/ in this sentence in different positions is listed in Table 1.

Word	Position	Duration (Sec)
n*aa*n	middle	0.15
*aa*ru	initial	0.16
varal*aa*maa	middle	0.15
varalaam*aa*	final	0.18

Table 1. Duration analysis of /aa/ occurring in various positions

Duration analysis as shown above is performed on a set of samples recorded from a native Tamil speaker. The duration information is tabulated and is stored as a look up table for future reference. Although loudness is not as important a phenomenon as pitch, the mismatch of loudness at concatenation points may introduce artifacts. The artifact may be in the form of an echo. This is caused due to the amplitude envelope mismatch. This is efficiently handled by matching the power at the concatenation point.

3.1.4. Consonant-Vowel Segmentation

It is observed that any change in the consonant part of a signal results in change of perception of the unit. Consonants must be kept intact. To this end, consonant and the vowel regions of the units must be segmented. Based on the point of origin in the vocal tract, consonants can be classified into co-articulated and non co-articulated. In the literature, segmentation of speech is applied mostly for speech

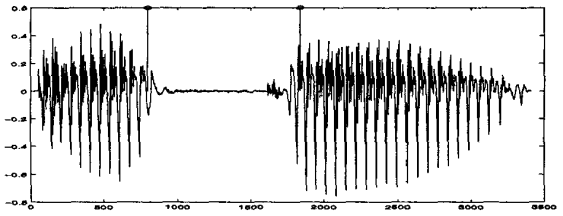

Fig. 2. Result of automated segmentation of non co-articulated consonant /aka/

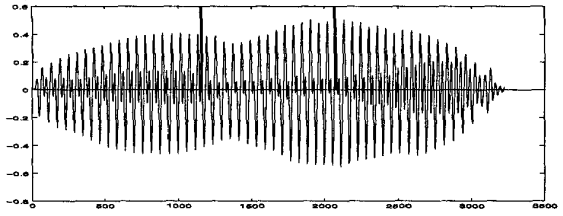

Fig. 3. Result of automated segmentation of co-articulated consonant /iyi/

recognition. It is an essential tool for building large corpora for training speech recognition systems. Manual segmentation is both time consuming as well as prone to error and inconsistencies [9][10]. We make use of a hierarchical method of segmentation to split the given phonemes into vowel and consonant regions [8]. In the primary level, the input phonemes are divided into two classes based on log of energy. The first class is segmented in the time domain and the other class, in the frequency domain. Difference in energy in successive frames is employed as a feature for segmentation in time domain approach and cepstral analysis is adopted in frequency domain. About 98% accuracy is achieved using this technique and heuristics are used to handle the remaining phonemes in the database. Figures 2 and 3 show the results of both co-articulated and non co-articulated consonant segmentation using the above technique.

3.1.5. Pitch Marking

Pitch marking [2] is essential as the waveforms are concatenated at the pitch marks. Also the method employed for pitch modification to improve naturality requires pitch marking. The DCT based Spectral auto-correlation function [6] is employed for estimation of the pitch. Pitch marking is performed only on the vowel part after getting the result from consonant-vowel segmentation as explained in section 3.1.4. . The pitch marked signal is shown in figure 4.

After the above specified operations, the data obtained viz, segmentation points, pitch marks and the wave data are stored in a customized file format to facilitate easy access during the online process. The duration and the intonation

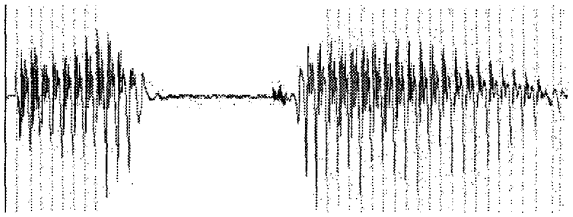

Fig. 4. Pitch marked non co-articulated consonant /aka/

information after studying from a native speaker's voiceare also stored in another data file.

3.2. ONLINE PROCESS

This involves the following phases: (1) Text analysis (2) Application of rules (3) Concatenation (4) Post processing

3.2.1. Text Analysis

Text analysis phase involves parsing the input text into a sequence of basic units of speech and tag them with appropriate emotions. The four basic punctuation marks viz, full stop, comma, exclamation mark and question mark are considered. The tagged information is efficiently made use of in post processing to make the synthesized speech more natural.

3.2.2. Application of Rules

Tamil grammar rules discussed in section 2 are applied on the parsed text. Prosodic information is applied and retagging is done on the already tagged text. This facilitates duration modeling as per observation in natural speech.

3.2.3. Concatenation

A very simple but working strategy has been adopted for concatenation. The difference in pitch period between the last pitch period of the first segment and the first pitch period of the following segment is made use of. If the difference between the periods is less than 5 samples (approx less than 0.75 msec), then the concatenation is performed at the lowest energy point within those pitch periods. If the difference is more than 5 samples, waveform interpolation is used for smoothing the spectrum. We generate new pitch periods using the algorithm in [3][7]. The algorithm is briefly explained as follows. The linear prediction residual is obtained from pitch synchronous frames by inverse filtering the speech signal. Then Discrete Cosine Transform (DCT) is applied on these pitch synchronous frames. Based on the desired factor of pitch modification, the dimension of the DCT vector is changed by truncation or zero padding,

	Naturalness			Intelligibility			Distortion		
	Good	Fair	Bad	Good	Fair	Bad	Low	Medium	High
Speech Output	75%	25%		90%	10%		95%	5%	

Table 2. Perceptive evaluation of ten listners for fifty different sentences

and then Inverse Discrete Cosine Transform is applied. This period modified residual signal is then forward filtered to obtain the pitch modified speech. The enhanced algorithm [11] of [3] and [7] has been used in Thirukkural.The same algorithm is used for pitch modification in post processing for introduction of emotions and for naturalizing the synthesized speech.

3.2.4. Post processing

In post processing we modify the pitch and the duration as per the tagging in the previous section. The factor of pitch modification for an exclamation, question, comma and full stop varies respectively from high to low. Pitch modification algorithm [3] explained in the above section is used for introducing naturalness.

4. IMPLEMENTATION

Thirukkural is designed to work on Windows 95, Windows 98 and Windows NT. It is designed using C++ and graphic user interface (GUI) is provided using Visual C++. The preliminary version of the software is available in JAVA as well. It supports TAB keyboard as recommended by Tamilnet99 standards. Phonetic Tamil keyboard has also been implemented. The feature for saving the synthesized wave file and for reading the existing text in TAB format is also provided.

5. PERCEPTUAL EVALUATION

To evaluate the overall performance, perceptive evaluation was carried out. Ten native Tamil speakers were asked to rate the synthesized speech in terms of intelligibility, naturalness and distortion on fifty different sentences. The result of this evaluation (listed in Table 2) showed excellent intelligibility, average naturalness and low distortion.

6. CONCLUSION

Thirukkural synthesizes intelligible Tamil speech. It has a male voice and can read input text, which is in TAB format. The techniques used are novel and the quality of speech is good. Attempts are being made to make it natural, add emotions, making it net enabled and also to provide good synthesis for alien words (like /fa/ in 'father'). The system is extendable to any language just by changing the language rules, intonation patterns and of course the database. A primitive system exists in Kannada, another south Indian language.

7. REFERENCES

[1] Douglas O'Shaughnessy (2000), Speech Communication, IEEE Press, New York.

[2] G L Jayavardhana Rama, A G Ramakrishnan, R Muralishankar and Vijay Venkatesh. "Thirukkural - A text to speech synthesis system". Proc. Tamil Internet 2001, Kuala Lumpur 2001, 92-97.

[3] R. Muralishankar and A. G. Ramakrishnan, "DCT Based Pitch Modification", Proc. of SPCOM '01, IISc, Bangalore, 2001, pp 114-117.

[4] K G Aparna, G L Jayavardhana Rama and A G Ramakrishnan. "Machine reading of Tamil Books - An aid for the blind". Proc. Biovision 2001, IISc, Bangalore, 173-177.

[5] Hunt, A. and Black, A."Unit selection in a concatenative speech synthesis system using a large speech database", Proceedings of ICASSP 96, Atlanta, Georgia, vol 1, pp 373-376.

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

[7] R Muralishankar, A G Ramakrishnan and P Prathibha, "Pitch modification using DCT in the Source Domain", Submitted to Journal on Speech Communication, after first revision.

[8] Jayavardhana Rama, R Muralishankar and A G Ramakrishnan, "Segmentation of Speech Units into Consonant and Vowel for Concatenative Speech Synthesis", Accepted for presentation in SPPRA 2002, GREECE.

[9] Ossama Essa, "Using Prosody in Automatic Segmentation of Speech", Proc. ACM 36th Annual Southeastern Conference, Atlanta, Georgia, April 1998.

[10] J. van Hemert,"Automatic segmentation of speech". IEEE, Trans. on Signal. Processing, 39(4), April 1991, 1008-1012.

[11] R Muralishankar and A G Ramakrishnan. Warped-LP Residual Resampling using DCT for Pitch Modification.Accepted for presentation in ICSLP 2002, Denver, Colorado, to be held during Sep16-20, 2002.