Development of Concatenative Syllable based Text to Speech Synthesis System for Tamil

B. Sudhakar
Department of Electrical Engineering
Annamalai University

R. Bensraj
Department of Electrical Engineering
Annamalai University

ABSTRACT
This paper addresses the problem of improving the intelligibility of the synthesized speech in Tamil TTS synthesis system. The human speech is artificially generated by Speech synthesis. The normal language text will be automatically converted into speech using Text-to-speech (TTS) system. This paper deals with a corpus-driven Tamil TTS system based on the concatenative synthesis approach. Concatenative speech synthesis involves the concatenation of the basic units to synthesize an intelligent, natural sounding speech. In this paper syllables are the basic unit of speech synthesis database and the modification of syllable pitch by time scale modification. The speech units are annotated with associated prosodic information about each unit, manually or automatically, based on an algorithm. An annotated speech corpus utilizes the clustering technique that provides way to select the suitable unit for concatenation, depends on the minimum total join cost of the speech unit**.** The entered text file is analyzed first, this syllabication is performed based on the linguistics rules and the syllables are stored separately. Then the syllable corresponding speech file is concatenated and the silence present in the concatenated speech is removed. After that discontinuities are minimized at syllable boundaries without degrading the quality. Smoothing at the concatenated syllable boundary is performed and changing the syllable pitches by time scale modification.

Keywords
Tamil TTS, Concatenative Speech Synthesis, Text to speech synthesis, Syllable based synthesis.

1. INTRODUCTION
Over the past years, there has been an immense development in Speech technologies. Among the applications of speech technology, the automatic speech production, which is referred to as TTS system is the most natural sounding technology. TTS synthesis is the process of converting ordinary orthographic text into speech signal which is indistinguishable from human speech [1-10]. It can be widely classified into front end and back end as shown in Fig.1.The conversion of natural language text to a structured linguistic representation is associated with front end. From the raw text this front end identifies a sequence of segments called target segments. These target segments have a different features estimated from the text. The back end is referred as the second part of the system which modifies these target segments into a speech waveform.

There are two main methods are used for speech production. These methods are format synthesis and concatenation synthesis is illustrated in [11]. The format synthesizer utilizes a simple model of speech generation and a set of rules to generate speech. While these systems can achieve enhanced intelligibility, their naturalness is typically low, since it is very tedious to perfectly describe the process of speech produced in a set of rules. The TTS has been the main research focus automatic speech production in Indian languages nowadays. Some of TTS systems for Indian languages like Hindi, Telugu, Tamil and Bengali have been developed using the unit selection and festival framework in [2] and [6]. Listeners are able to clearly perceive the message

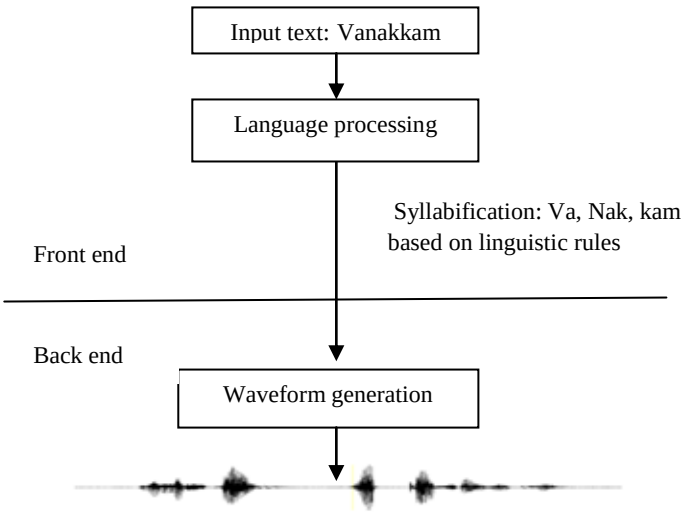

Fig. 1. Parts of speech synthesis system

with little attention, and act on synthesized speech of a command correctly and without perceptible delay in noisy environments. Although many TTS approaches, the intelligibility, naturalness, comprehensibility, and recall ability of synthesized speech is not good enough to be widely accepted by users. There is still considerable rule for further improvement of performance of the text-to speech production system.

This paper proposed corpus driven TTS system. In this paper the concatenative-based approach is used to produce desired speech through pre-recorded speech waveforms. Over past decades, this proposal was very complicate to implement because of limitation of computer memory. With the advancements in computer hardware and memory, a large quantity of speech corpus can be stored and utilized to produce high quality speech signal for a given text. Thus, the synthesized speech preserves the naturalness and intelligibility. Here the given input text is analyzed first. Based on the linguistics rules, syllabication is performed. The

syllables are called basic speech units. The repository of these units is created with its prosodic information.

The pitch value for syllable is changed by performing time scale modification. During the synthesis, these units are selected and concatenated with lowest join cost. After performing concatenation the waveform is smoothened at concatenation joints using LPC. The rest of this paper is organized as follows. In section 2 discussed about the concatenative speech synthesis system and syllabification rules for Tamil. In section 3 described the proposed Tamil speech to text synthesis system. Section 4 discussed about the quality test. Finally Section 5 provided synthesized speech waveforms and conclusions.

2. CONCATENATIVE SPEECH SYNTHESIS

Concatenative speech synthesis utilizes phones, diphones, syllables, words and sentences as basic units. Based on selecting these units from the database speech are synthesized, called as a speech corpus. Many researches have been made, selecting each separate unit as the basic unit. When phones are selected as basic units, for Indian languages the size of the database will be less than 50 units. Database may be small, but phones gives very poor co-articulation data's across neighboring units, thus falling to model the dynamics of speech sounds. Diphones and triphones as basic units, it will minimize the discontinuities at the concatenation points and captures the co-articulation effects. But a single example of each diphone is not enough to generate precious quality speech. So this paper presented a syllable as a basic unit. Indian languages are syllable centered, where pronunciations' are based on syllables. For Indian languages intelligible speech synthesis a syllable can be the best unit.

The general form of Indian language syllable is C*VC*, where C is a consonant, V is vowel and C* indicates the presence of 0 or more consonants. There are 18 consonants and 12 vowels in Tamil languages. There are defined set of syllabification rules formed by researchers, to generate computationally reasonable syllables. Some of the rules used to perform grapheme to syllable conversion [12] are:

- Nucleus can be Vowel (V) or Consonant (C)
- If onset is C then nucleus is V to yield a syllable of type CV
- Coda can be empty of C
- If character after CV pattern are of type CV then the syllables are split as CV and CV
- If the CV pattern if followed by CCV then syllables are split as CVC and CV
- If CV pattern is followed by CCCV then the syllables are split as CVCC and CV
- If the VC pattern is followed the V then the syllables are split as V and CV
- If the VC pattern is followed by CVC then the syllables are split as VC and CVC

The following new rules have been implemented in this paper to implement grapheme to syllable conversion

- If character after CV pattern are of type CV then the syllables are split as CVCV
- Similarly If character after CV pattern are of type CVCV then the syllables are split as CVCVCV
- If the CV pattern if followed by CVC then syllables are split as CVCVC
- If the CV pattern if followed by CCV then syllables are split as CVCCV

This paper proposed the following recommended combinations to achieve the best acceptable for synthesis is:

- Monosyllables at the beginning of a word and bisyllables at the end.
- Bisyllables at the beginning of a word and monosyllables at the end.
- Monosyllables at the beginning and trisyllables at the end of a word.
- Trisyllables at the beginning and monosyllables at the end of a word.

3. PROPOSED TAMIL TTS SYNTHESIS SYSTEM

3.1 Text analysis

To implement the proposed TTS system, the MATLAB 2012 has been used. In text analysis, first stage is text normalization then performs removing of punctuations such as double quotes, full stop, and comma. A pure sentence is synthesized at the end of text analysis. Then all the abbreviations present in the input text are expanded and also unwanted punctuation like (:,; ‘ $ `) etc. are removed to avoid confusion and not to give any disturbance in the naturalness of the speech. The next step in the text normalization is normalizing non-standard words like abbreviations and numbers. The next stage in the text analysis is sentence splitting. In this stage, the given paragraph will be spitted as sentences. From these sentences, words are separated out. The last stage is Romanization which is the representation of written words with a roman alphabet. In this system Romanized form of Tamil word/syllables are generated.

3.2. Speech corpus

Building a speech corpus for Indian languages is a difficult task than that of English speech corpus. Prosodic information such as pitch, duration and intonation prediction has to be done in the corpus development stage itself, some more information has to be specified with the basic speech units after storing them in the corpus. The problem such as mispronunciation, un transcribed speech units, phrase boundary detection, pronunciation variants are to be identified and addressed. For corpus creation we selected one person for recoding these basic units, who has uniform characteristics of speaking, pitch rate and energy profile and developed speech corpus in [2]. The digitized speech signal with sampling rate of 16 KHz and 16-bit resolution (Pulse Code Modulation uncompressed data format) proposed in [2]. The speech wave files are saved according to the requirement. The speech wave files corresponding to the Tamil words are named according to their corresponding Romanized names. The words collected comprises dictionary words, commonly used words, Tamil newspapers and story books, also different domain such as sports, news, literature and education for building unrestricted TTS initiated in [2].

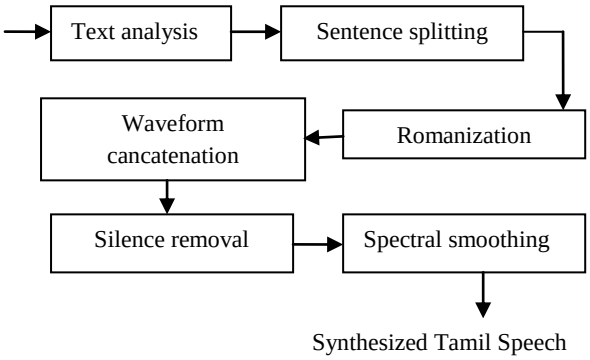

Fig. 2 Block diagram of proposed Tamil text to speech synthesis system

3.3 Waveform concatenation

In the final stage of the concatenation process, the required syllables are retrieved from the corpus based on the text analysis and arranged to produce the speech. Then all the arranged speech units are concatenated using a concatenation algorithm. The main problem in concatenation process is that there will be glitches in the joint. These are removed in the waveform smoothening stage. The concatenation process combines all the speech files which are given as an output of the unit selection process and then making in to a single speech file.

3.4 Spectral smoothing

The time scale modification is carried out for each syllable to produce individual smoothness for syllable in Tamil TTS. The time scale modification is used to change the pitch value for Tamil syllable. Praat software is used to calculate the Duration value for each syllable [13]. Smoothing at concatenation joints are performed using Linear Predictive Coding (LPC). The LPC is used for representing the spectral envelope of a digital signal of speech using the information of a linear predictive model. It is one of the most powerful methods for encoding enhanced quality speech at a low bit rate and gives extremely accurate estimates of speech features in [14],[15].Now we are getting the improved quality speech for the given input text. It can be played and stopped any where needed. The main aim of the proposed scheme is to achieve good naturalness in output speech. Fig.3 shows the smoothened output waveform.

3. QUALITY TEST

For developing a Tamil TTS, we have considered 700 sentences for recording the speech corpus. These are selected from various domains from newspaper, Wikipedia, news broadcast and story books. After formulating the data, speech corpus will be recorded in a studio environment with a suitable trained speaker. Recorded 700 sentences of speech corpus is classified into two parts: (i) part-1 contains 600 sentences for building the training corpus. (ii) part-2 contains 100 sentences for evaluating the established Tamil TTS system. Five speakers' voices are selected for constructing prototype TTS systems. From this proposed work the following performance are inferred that the speech of five speakers with respect to (i) the quality of the synthesized speech (ii) Variations in natural prosody and (iii) the perceptual distortion with respect to prosodic and spectral modifications.

Evaluation of quality of the synthesized speech is carried out by subjective measures. An intelligibility and naturalness are estimated from the listening tests. Tests are conducted with 25 research scholars in the age group of 23–35 years. The subjects have sufficient speech knowledge for proper assessment of the speech signals, as all of them have taken a full semester course on speech technology. Speech utterances corresponding to the test sets are synthesized using the developed Tamil TTS system. Each of the subjects was given a pilot test about perception of speech signals by playing the original speech samples of the test files. Once they are comfortable with judging, they are allowed to take the tests. The tests are conducted in the laboratory environment by playing the speech signals through headphones. In the test, the subjects were asked to judge the distortion and quality of the speech. Subjects are asked to assess the quality and distortion on a 5-point scale for each of the sentences. The 5-point scale for representing the quality of speech and the distortion level is given in Table 1.

For evaluating the quality of synthesized speech generated from the developed TTS system, there are three sets of test utterances are considered. Each set consists of 20 sentences. set-1: All the words are available from the training data, but the entire word sequences are not present in the training data. Set-2: 50% of the words available in training corpus. Set-3: none of the words are available in the training corpus. Table 2 shows the MOS scores for the three test sets. From the table 2 it is observed that, MOS for set-1 is more compared set-2 and set-3 as all the words of sentences in set-1 is present in database. So it provides the better performance compared to other two sets.

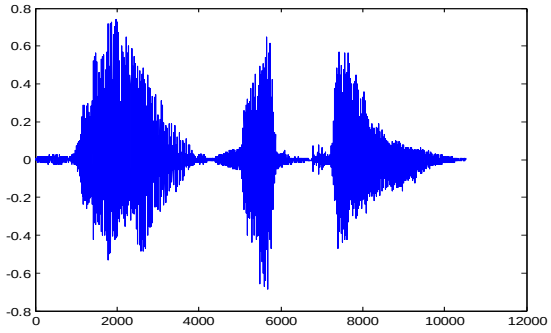

Fig.3. Resulting concatenated waveform after performing silence removal

Table 1. Instructions to evaluators

Score	Subjective perception
1	Poor speech, with distortion and very low intelligibility
2	Poor speech with distortion and intelligible
3	Good speech with less distortion and intelligibility
4	Very good speech quality with less naturalness
5	As good as natural speech

Table 2. Mean opinion scores for three sets

Test Set	MOS
Set – I	4.01
Set – II	3.25
Set - III	2.97

5. RESULTS AND CONCLUSION

In this proposed word, a speech synthesis system has been designed and implemented for Tamil language. A database has been created from various domain words and syllables. Syllable pitch modification is performed based on time scale modification. The speech files present in the corpus are recorded and stored in PCM format in order to retain the naturalness of the synthesized speech. The given text is analyzed and syllabication is performed based on the rules specified. The desired speech is produced by concatenative speech synthesis approach such that spectral discontinuities are minimized at unit boundaries. It is inferred that the produced synthesized speech is preserving naturalness and good quality based the subjective quality test results. The final output speech file is stored in the specified location in the system for further analysis.

6. REFERENCES

[1] Marian Macchi, Bellcore. 1998. Issues in text-to-speech synthesis, In Proc. IEEE International Joint Symposia on Intelligence and Systems, pp.318-325.

[2] N.P. Narendra, K. Sreenivasa Rao ,Krishnendu Ghosh, Ramu Reddy Vempada, Sudhamay Maity. 2011. Development of syllable-based text to speech synthesis system in Bengali, International journal of speech technology, pp.176-181.

[3] C. Pornpanomchai, N. Soontharanont, C. Langla, N. Wongsawat. 2011. A dictionary-based approach for Thai text to speech (TTTS), In Proc. Third Int. Conferenceon Measuring Technology and Mechatronics Automation, vol. 1, pp.40-43.

[4] M. N. Rao, S. Thomas, T. Nagarajan, and H. A. Murthy. 2005. Text-to-speech synthesis using syllable like units, In National Conference on Communication, IIT Kharagpur, pp. 227–280.

[5] D. H. Klatt. 1987. Review of text-to-speech conversion for English, The Journal of the Acoustical Society of America, pp.737–793.

[6] M. Sreekanth, and A.G. Ramakrishnan. 2007. Festival based maiden tts system for Tamil Language, In Proc. 3rd Language and Technology Conference, Poznan, Poland, October, pp. 187–191.

[7] S. P. Kishore and A. W. Black. 2003. Unit size in unit selection speech synthesis. In Proc. Eurospeech.

[8] N.S. Krishna, P.P. Talukdar, K. Bali, A.G. Ramakrishnan. 2004. Duration modeling for Hindi text-to-speech synthesis system, In Proc. of Intenational Conference on Spoken Language Processing (ICSLP'04), Korea.

[9] A. Hunt, and A. Black. 1997. Unit selection in a concatenative speech synthesis system using a large speech database, In Proc. of IEEE Int. Conference Acoustic, Speech, and Signal processing, vol. 1, pp. 373–376.

[10] Robert J. Utama, Ann K. Syrdal, and Alistair Conkie. 2006. Six approaches to limited domain concatenative speech synthesis, INTER SPEECH, ICSLP.

[11] S.D. Shirbahadurkar, D.S. Bormane, R.L. Kazi. 2010. Subjective and spectrogram analysis of speech synthesizer for Marathi tts using concatenative synthesis, Recent Trends in Information, Telecommunication and Computing (ITC).

[12] S Saraswathi , T V Geetha. 2010. Design of language models at various phases of Tamil speech recognition system, International Journal of Engineering, Science and Technology, Vol. 2, No. 5, pp. 244-257.

[13] K. P. Mohanan, T. Mohanan. 1987. Lexical phonology of the consonant system in Malayalam, Linguistic Inquiry, The MIT Press, volume 15.

[14] K. Panchapagesan, P.P Talukdar, N.S. Krishna, K.Bali and A.G. Ramakrishnan. 2004. Hindi text normalization, Fifth International Conference on Knowledge Based Computer Systems (KBCS), Hyderabad, India.

[15] T.Charoenporn, A.Chotimongkol, V.Sornlertlamvanich. 1999. Automatic romanization for Thai, In *Proc.* of the 2nd Int.workshop on East-Asian Language Resources and Evaluation.