IJETAE
EXPLORING RESEARCH AND INNOVATIONS

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459 (Online), An ISO 9001:2008 Certified Journal, Volume 3, Special Issue 1, January 2013)

International Conference on Information Systems and Computing (ICISC-2013), INDIA.

TEXT TO SPEECH SYNTHESIS SYSTEM FOR TAMIL

J.Sangeetha[1], S. Jothilakshmi[2], S.Sindhuja[3], V.Ramalingam[4]

Department of Computer Science & Engineering, Annamalai University, Chidambaram, India.

jayasangita@yahoo.com, sindhusm77@gmail.com, aucsevr@gmail.com, jothi.sekar@gmail.com

Abstract

In a text-to-speech system, spoken utterances are automatically produced from text. In this paper, we present a corpus-driven Tamil text-to-speech (TTS) system based on the concatenative synthesis approach. The most important qualities of a synthesized speech are naturalness and intelligibility. In this system, words and syllables are used as the basic units for synthesis. Our corpus consists of speech waveforms that are collected for most frequently used words in different domains. The speaker is selected through subjective and objective evaluation of natural and synthesized waveform. The proposed system provides utility to save the synthesized output. The output generated by the proposed Tamil text-to-speech synthesis system resembles natural human voice. Our text to speech reader software converts a Tamil text to speech wav file that has high rates of intelligibility and comprehensibility.

***Keywords*-- Text-to-Speech, Syllabification, Concatenation, Speech Synthesis, Text Normalization.**

I. INTRODUCTION

Over the past years, there has been a great development in speech understanding and synthesis technology. The voice user interface (VUI) plays an important role in human-machine communication applications such as computer systems, mobile multimedia, online ticket information, market information, customer services, personal banking information, voice-enabled equipment maintenance devices, and paperless tasks. Most of these voice-enabled applications have imparted huge financial benefits for the multimedia industries. Among the applications of speech technology, the automatic speech production, which is referred to as text-to-speech (TTS) system is the most natural-sounding technology. The text-to-speech (TTS) system will convert ordinary orthographic text into acoustic signal which is indistinguishable from human speech [1]-[10].Today‗s interest is high quality speech application combined with computer resources. Text-to-speech synthesis system can be useful for several multimedia applications. For developing a natural human machine interface, the TTS system can be used as a way to communicate back through human voice. The TTS can be a voice for those people who cannot speak. The TTS system can be used to read text from emails, SMSs, web pages, news, articles, blogs, and Microsoft office tools and so on. In such reading applications, the TTS technology can reduce the eye-strain. The TTS system can be useful for disabled person to make effective communications. The existing TTS systems can be broadly classified into three groups: i) articulatory synthesis;ii) formant synthesis; iii) concatenative synthesis [11], [12]. In the past decades, the TTS has been focused on automatic speech production in Indian languages.

Some of TTS systems for Indian languages like Hindi, Telugu, Tamil and Bengali have been developed using the unit selection and festival framework [2], [6]. In literature, each approach has its own purposes, strengths, and limitations. In practice, listener should be able to understand language information of the user textual information in generated synthesized speech waveform. In most of multimedia applications, listeners demand high quality of synthesized speech compared with natural speech. Generally speaking, the intelligibility and comprehensibility of synthesized speech should be relatively good in the naturalistic environments. Furthermore, listeners are able to clearly perceive the message 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 room for further improvement of performance of the text-to speech production system. In this paper, we propose corpus driven Tamil text-to-speech system. We use concatenative-based approach to synthesis desired speech through pre-recorded speech waveforms. Over past decades, this approach was very difficult to implement because of limitations of computer memory. With the advancements in computer hardware and memory, a large amount of speech corpus can be stored and used to produce high quality speech waveforms for a given text. Thus, the synthesized speech preserves the naturalness and intelligibility. In this work, we applied concatenation approach at word level and syllable level. The quality of synthesized speech via concatenative approach is very close to natural speech.

The rest of this paper is organized as follow. In Section 2, we discuss Tamil phonology briefly. In Section 3, we discuss the detailed descriptions of concatenative speech synthesis. In section 4, we describe the proposed Tamil text-to speech system. Finally, we provide synthesized speech waveforms and conclude in Section 5.

II. Tamil Phonology

Tamil is one among the Dravidian languages in India. Tamil is the official language of Kerala state and the Union Territories of Lakshadweep and Pondicherry. Tamil language contains 2500 unique phonemes. Difficulties in developing Tamil TTS include understanding Tamil phonetics, database creation of Tamil language, syllable level concatenation, complexity of the language etc. There are 18 consonants and 12 vowels in Tamil language [13].

III. Concatenative Speech Synthesis

Concatenative speech synthesis uses phones, diphones, syllables, words and sentences as basic units.

Speech is synthesized based on selecting these units from the database, 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, the size of the database will be less than 50 units for Indian languages. Database may be small, but phones provide very less co-articulation information across adjacent units, thus failing 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 produce good quality speech. So we are selecting syllable as a basic unit. Indian languages are syllable centered, where pronunciations are based on syllables. 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 defined set of syllabification rules formed by researchers, to produce computationally reasonable syllables. Some of the rules used to perform grapheme to syllable conversions [17] are:

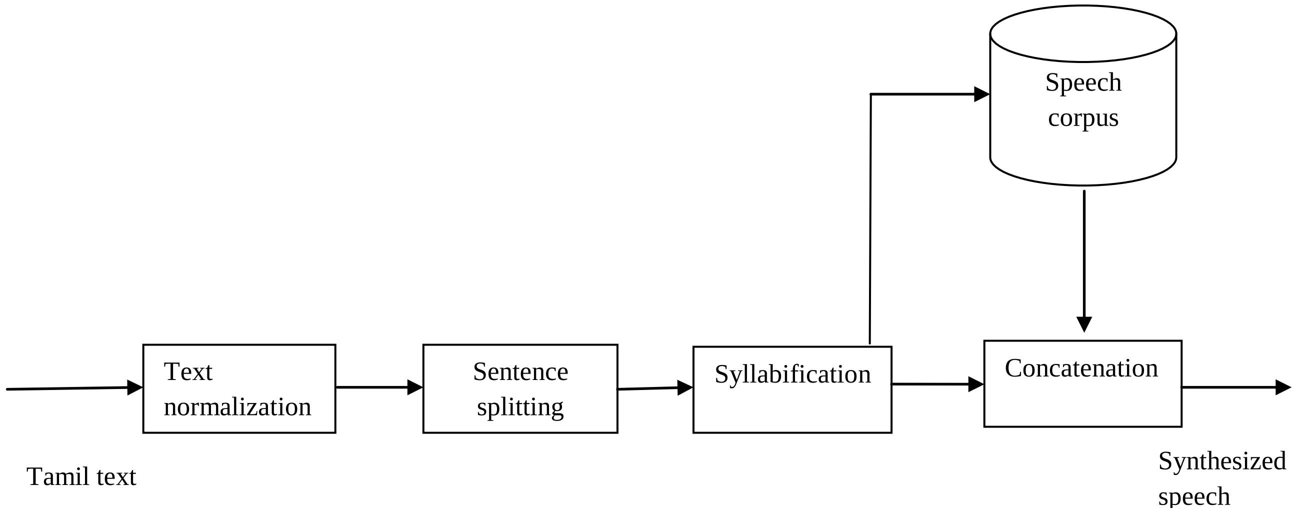

Fig. 1.Block Diagram of Tamil Text-To-Speech Synthesis System

- 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 by 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

These rules can be generalized to any syllable centric language. The text is pre-processed to remove any punctuations. Numerals1,2,3 and 4 in the algorithm represent the position of the alphabet in the text to be segmented. Algorithm for these rules are:

```
•Check the first character of the word:
•If the first character is a V then
the second character is a C ; Check the third character:
   – If the third character is a C; Check fourth character:
         *If 3rd and 4th characters are equal;
            VCC(123)is the syllable
         ∗ else; VC (12) is the syllable
– If the third character is a V ; Then V (1) is the syllable
• If the first character is a C then
Check for second character:
– Second character is a V
   The third character has to be a C; Check the 4th character:
        ∗If the fourth character is a C ;Check for
           5thcharacter:
           5th character is a V; CV C (123) is the
           syllable5th character is a C or a word end; CV
           CC(1234) is the syllable
      ∗ If the fourth character is a V ;
         CV (12) is the syllable
– Second character is a C; we assume that the 3rd
     character has to be a vowel, and subsequently the
     4th character has to be a C Check for 5th character:

     ∗If the 5th character is a C or a word end; Then
        CCVC (1234) is the syllable.
∗If the 5th character is a V;
   Then CCV(123) is the syllable.
```

Text syllabication example using the above mentioned algorithm:

a	**var**	**sey**	**di**	**yA**	**lar**
V	**CVC**	**CVC**	**CV**	**CV**	**CVC**

’|’ represents a word boundary. It should be noted that each word is syllabified separately. After a syllable is identified from a word, the remaining part of the word is processed again by the algorithm. The text syllabification algorithm gives units comparable to the units given by group delay based segmentation. The two units can be made equivalent by using some specific language or domain rules. For example, it was observed that almost always ”diru” was pronounced as a single unit ”diR”. Once the units are comparable or equivalent the segmented text can annotate the speech syllables. These syllabified texts can also be used to analyse syllable structure in the language like frequently occurring syllables or the syllables that can start or end a sentence. Syllable based N-gram language models can be built using these rules to segment large amount of text.

IV. Proposed Tamil Text-To-Speech Synthesis System

In **Fig. 1** illustrates the steps involved for the conversion of Tamil text to speech.

4.1. Text Normalization

In this stage, we perform remove punctuations such as double quotes, full stop, comma and all. Then we will get pure sentence. We need to know that the sentence ends after a full stop (.) and not between abbreviations. It is somewhat easy to tokenize a word with help of full stop as most of the sentences will be ending with full stop. But there are some other cases where it ends with semicolon or some other punctuation like previous case. This problem can be solved by expanding the abbreviation and removing the unwanted punctuation.

All the Tamil abbreviations cannot be expanded, because some mostly used abbreviations are stored in a separate database. When certain abbreviation comes in the text, then it will search the database for that abbreviation. If that abbreviation is present the system will replace the text, if not it will be leaving the original text as it is. It is difficult to add all the abbreviations in the database, so most commonly used abbreviations are used. The unwanted punctuation like (: , ; " ' ` $) etc. are to be removed from the given paragraph to avoid confusion and not to give any disturbance in the naturalness of the speech. Each and every text in the input should be assigned some sound file for the concatenation.

The second step in text normalization is normalizing non-standard words. Non standard words are tokens like numbers or abbreviations, which need to be expanded into sequences of Tamil words before they can be pronounced. For example the number 1900 can be spoken in at least three different ways, depending on the context.

¬Â¢ÃòÐ ¦¾¡Ç¡Â¢Ãõ
Àò¦¾¡ýÀÐ áÚ
´ýÚ ´ýÀÐ âƒ¢Âõ âƒ¢Âõ

To solve this problem we have to add number system into TTS. It will be coming under future work. The algorithm will cancel all the numbers coming in text, and then it will become normal text without numbers or any extra punctuation.

4.2.Sentence Splitting

In this stage, the given paragraph will be splitted as sentences. Separating out sentences can also be done in parallel through Graphical Processing Unit computing. From these sentences, words are separated out. Example is given below

¿¡ý þíÌ ¿¢ü¸¢§Èý --> (¿¡ý . þíÌ . ¿¢üì¸¢§Èý)

The written sentence can be segmented easily by using whitespace as delimiter,

1921-->
(¬Â¢ÈòÐ.¦¾¡Ç¡Â¢ÈòÐ.þÕÀòÐ.´ýÚ) «øÄÐ (´ýÚ.´ýÀÐ.þÃñÎ.´ýÚ)

There are numerous cases where the conventional delimiter segmentation approach fails. The classification and segmentation cannot be done one after the other. The first step to perform a provisional segmentation into potential written form is called token and then examining each token in turn resolves the ambiguity. This process is called tokenization; the step which generates the words from the token is called text analysis [1].

4.3.Speech corpus

Text-to-speech system based on concatenative synthesis needs well arranged speech corpus. The quality of synthesized speech waveform depends up on the number of realization of various units present in the speech corpus. A good quality microphone should be used to avoid noise in speech wav file. In text-to-speech, the accuracy of the system is calculated in the ways of naturalness and intelligibility of the synthesized speech. In this work, we collected speech wav files for 5000 Tamil words and syllables.

We selected one person for recoding these words, who has uniform characteristics of speaking, pitch-rate and energy profile, and developed speech corpus [2]. Each sound file is unique. Speech corpus collected includes text from dictionary words, commonly used words, words from Tamil newspapers and story books, and covers different domain such as sports, news, literature, education etc. We have analyzed the speech with respect to (1) the quality of the synthesized speech (2) variations in natural prosody and (3) the perceptual distortion with respect to prosodic and spectral modifications.

4.4.Concatenation

The final stage is the concatenation process. All the arranged speech units are concatenated using a concatenation algorithm.

The concatenation of speech files is done in matlab.

4.5. Speech synthesis

In speech synthesis we are utilizing two approaches, the first one is word level synthesis that means all the words that are present in the input text is already in the speech corpus so synthesized output naturalness is very high. Second, when input word is not present in the database we synthesis the word using syllable level concatenation. In this case naturalness will bec omparatively less than word level synthesis. For Example,

¸¡¨Ä Å½ì¸õ ÜÚ¸¢È¡ý

The system first removes"." from the input text then it check the spaces between words and break the given input into different lines. So the input sentence become 3 words

¸¡¨Ä
Å½ì¸õ
ÜÚ¸¢È¡ý

Then the matlab program searches the normalized database to find whether the word is present or not using the help of mapping file.

V. QUALITY TEST

Voice quality testing is performed using subjective test. In subjective tests, human listeners hear and rank the quality of processed voice files according to a certain scale. The most common scale is called MOS (Mean Opinion Score) [16] and is composed of five scores of subjective quality, 1-Bad, 2-Poor, 3- Fair, 4-Good, 5-Excellent. The MOS score of a certain vocoder is the average of all the ranks voted by different listeners of the different voice file used in the experiment.

Here tests are conducted with five students with the age group of 20-26 years. The tests were conducted in the laboratory environment by playing the speech signals through headphones. In this test, these five students were asked to audio perception for the three synthesized sentences. Then they were asked to judge the distortion and quality of the speech. The evaluation of Tamil TTS is shown in **Table 1.**

Sentence-1 score is obtained by making 5 people to speak and evaluate the speech output of the test sentence with words present in the speech corpus. To obtain the score for Sentence-2, the test sentence is made with words present in the speech corpus and those which are not in the speech corpus. The score for Scentence-3 is obtained by making the test sentence with words not in the speech corpus.

Table 1.
Subjective test results.

Test Set	MOS
Sentence - I	4.00
Sentence - II	3.50
Sentence - III	3.00

VI. Result And Conclusion

In this paper, a speech synthesis system has been designed and implemented for Tamil Language. A database has been created from the various domain words and syllables. 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 syllabified based on the syllable segmentation rules. The desired speech is produced by the concatenative speech synthesis approach. **Fig. 2**shows the concatenated synthetic speech signal for Tamil phrase. Thus we can conclude that the produced synthesized speech is preserving naturalness and good quality based on the subjective quality test results.

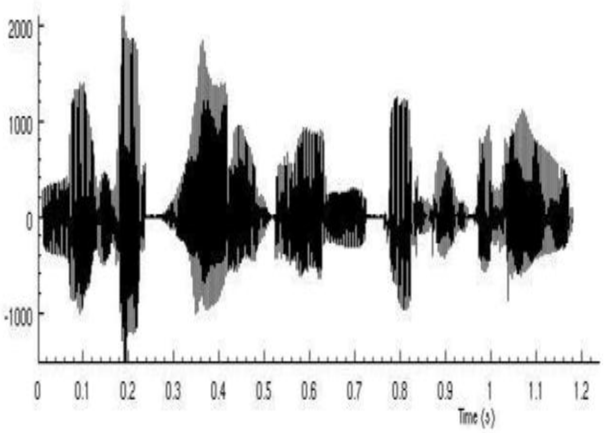

Fig.2: concatenated synthetic speech signal for Tamil phrase" ¸¡¨Ä Å½ì¸õ ÜÚ¸¢È¡ý"

REFERENCES

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

[2] N.P. Narendra, K. SreenivasaRao ,Krishnendu Ghosh, Ramu ReddyVempada, SudhamayMaity. Development of syllable-based text to speech synthesis system in Bengali.

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

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

[5] D. H. Klatt. Review of text-to-speech conversion forEnglish. The Journal of the Acoustical Society of America,82, 737 – 793, 1987

[6] M. Sreekanth, & A.G. Ramakrishnan. Festival basedmaiden tts system for Tamil language. InProc. 3rd language and technology conf., Poznan, Poland, October pp. 187 – 191, 2007.

[7] S. P. Kishore and A. W. Black. Unit size in unit selectionspeech synthesis.In Proc. Eurospeech2003, Sept. 2003.

[8] N.S. Krishna, P.P. Talukdar, K. Bali, A.G. Ramakrishnan. Duration modeling for Hindi text-to-speech synthesis system. InProc. Of Int. Conference on Spoken Language processing ICSLP" 04, Korea, 2004

[9] A. Hunt, & A. Black. Unit selection in a concatenative speech synthesis system using a large speech database. InProc. of IEEE int. Conference acoust, speech, and signal processing, vol. 1, 2004

[10] N. Sridhar Krishna, HemaA. Murthy and Timothy A.Gonsalves,. Text-to-Speech (TTS) in Indian Languages.Int. Conference on Natural Language Processing, ICON-2002, Mumbai, pp. 317.326, 2002.

[11] Robert J. Utama, Ann K. Syrdal, Alistair Conkie. Six approaches to limited domain concatenativespeech synthesis. INTERSPEECH, ICSLP, 2006

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

[13] K. P. Mohanan, T. Mohanan..Lexical phonology of the consonant system in Malayalam". Linguistic Inquiry The MIT Press, volume 15, 1984.

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

[15] T.Charoenporn, Sornlertlamvanich,. An Automatic romanization for Thai. InProc. of the 2nd Int. Workshop on East-Asian Language Resources and Evaluation,1999

[16] M. Cernak, M. Rusko,. An evaluation of synthetic speech using the PESQ measure. InProc. Forum Acusticum, Budapest, 2725-2728,2005

[17] Lakshmi A, Hema A Murthy. A Syllable Based Continuous Speech Recognizer for Tamil. In Proc. of the 2nd Int. Workshop on East-Asian Language Resources and Evaluation,2009.