INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH & DEVELOPMENT

ISSN: 2278 – 0211 (Online)

Dravidian – Tamil Tts For Interactive Voice Response System

S. Jothilakshmi
Department of Computer Science & Engineering, Annamalai University, Chidambaram, India
S.Sindhuja
Department of Computer Science & Engineering, Annamalai University, Chidambaram, India
V. Ramilingam
Department of Computer Science & Engineering, Annamalai University, Chidambaram, India

Abstract:

Speech synthesis is the artificial production of human speech. A Text-To-Speech (TTS) system converts normal language text into speech automatically. The main application of Tamil TTS is aid to vocally handicapped guys. In this paper, we present a Dravidian - Tamil text-to-speech system based on the concatenative synthesis approach. Concatenative speech synthesis involves the concatenation of the basic units to synthesize an intelligent, natural sounding speech. The database consisting of the units along with their annotated information is called as the annotated speech corpus. Here, the entered text file analyzed first, syllabication is performed based on the syllabification rules and the syllables are stored separately. Then the corresponding speech file for the syllables are retrieved, concatenated and the silence present in the concatenated speech is removed and the synthesized speech is produced with good quality. Finally, the Smoothening with Optimal Coupling technique is applied for smooth the transition between concatenated speech segments in order to produce continuous output and the system resembles natural human voice. The most important qualities of a synthesized speech are naturalness and intelligibility.

Keywords: *Syllable segmentation, speech concatenation, Silence reduction, Smoothening with Optimal Coupling.*

1.Introduction

Speech and spoken words have always played a big role in the individual and collective lives of the people [1]. Wars have been won, peace agreements have been made because of the magical words of a few who knew how to give life to their words. Speech represents the spoken form of a language and is also one of the important means of communication. Over the past few decades, many researchers have been done in the field of converting text to speech. This research has resulted in important advances with many systems being able to generate a close to a real natural sound. These advances in speech synthesis also pave the way for many new speech related applications. The function of text-to-speech (TTS) system is to convert an arbitrary text to a spoken waveform. This generally involves two steps, *i.e.*, text processing and speech generation. Text processing is used to convert the given text to a sequence of synthesis units while speech generation is generation of an acoustic wave form corresponding to each of these units in the sequence [2]. Fig. 1 shows the general structure for text to speech synthesis system.

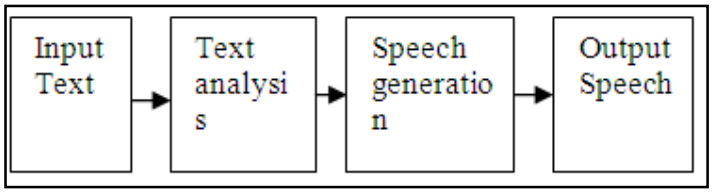

Figure 1: Text to Speech system

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 [3]-[8]. 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 existing TTS systems can be broadly classified into three groups: i) articulatory synthesis; ii) formant synthesis; iii) concatenative synthesis [9], [10]. Developing speech synthesis system is a complicated process and it also have some challenges [10].

Development of TTS systems require knowledge about human speech production and about languages being developed. The actual implementation of a fully functional system requires good software skills.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. In Section 2, we discuss Tamil phonology briefly. In Section 3, we discuss the detailed descriptions of Tamil TTS system. In section 4, the results are analyzed. Finally, we provide synthesized speech waveforms and conclude in Section 5.

2.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 [7].

3.Proposed Work

In Fig. 2 illustrates the steps involved for conversion of Tamil text to speech.

3.1.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. Examples is given below

- ¿¡ý Ş¸¡Å¢ÖìÌ ¦°ø¸¢ŞÈý.-->
  (¿¡ý . Ş¸¡Å¢ÖìÌ. ¦°ø¸¢ŞÈý)

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

- «¨ÉÅÕìÌõ Å½ì¸õ --> «¨ÉÅÕìÌõ
  - Å½ì¸õ

The written sentences can be segmented easily by using whitespace or '-' as delimiter.

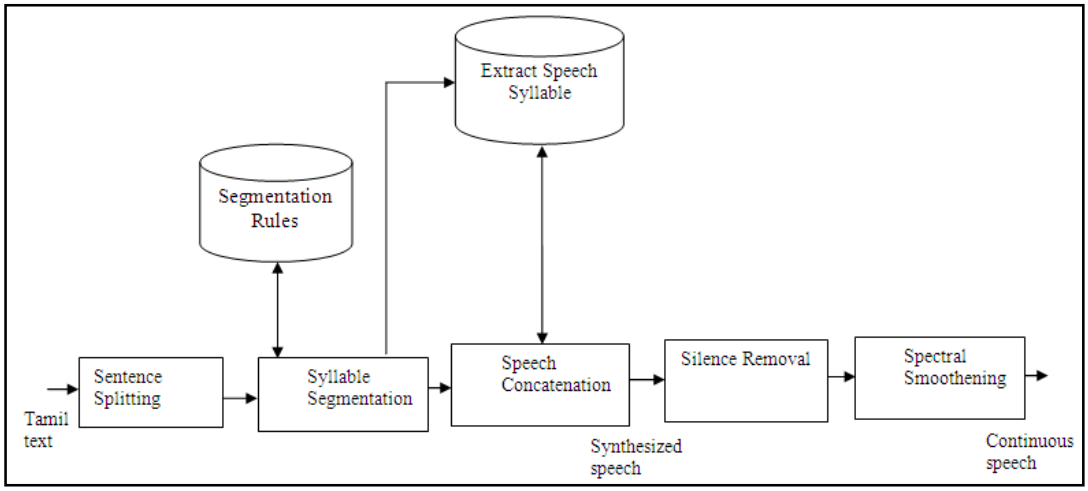

Figure 2: Overall Steps for Tamil TTS System

3.2.Syllable Segmentation

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.

3.2.1.Syllabification Rules

- 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 or CVCV
- If CV pattern is followed by CVC then the syllable is CVCVC
- If the CV pattern if followed by CCVCV then syllable is CCVCV
- If character after VC pattern are of type CV then the syllable is VCCV
- If the CV pattern if followed by CCV then syllables are split as CVC 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
- If the CV pattern is followed by CCVC then the syllable is CVCCVC

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.

3.2.1.1.Algorithm for Text Syllabification

• Check the first character of the word:

• If the first character is a V then the second character is a C ;

– If the third character is a C; Check fourth character:

* If 3rd char is C then fourth character is V; then VCCV

* else; VC is the syllable.

– If the third character is a V ; Then V is the syllable and CV is another syllable or VCV 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, CVC is a syllable.

* If the fourth character is a C; Check for 5thcharacter:

*5th character is a V; CVCCV is the syllable is the syllable

* If the 6th character is C and 7th is V then CVCCVCV is the syllable.

* If the fourth character is a V ; CV is the syllable or CVCV 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 examples using the above mentioned algorithm:

VC – ¬ñ

VCV – þ/¨º

VCCV –«õ/Á¡

CV – â

CVC – ¸ñ

CVCV – â/¨ƒ

CVCVC – ¾/Á¢ú

CVCVCV – ¯/¾/Å¢

CVCVCVC – Á¡/½/Åý

CVCCVCV – ¿ý/È¡¸

'/' represents a syllable boundary. 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.

3.3.Speech Concatenation

Concatenative speech synthesis is based on the concatenation (or stringing together) of segments of recorded speech. Generally, concatenative synthesis produces the most natural-sounding synthesized speech. However, differences between natural variations in speech and the nature of the automated techniques for segmenting the waveforms sometimes result in audible glitches in the output.

3.4.Speech Synthesis

In speech synthesis we are utilizing two simple approach, the first one is 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 be comparatively less than word level synthesis. For Example,

Å½ì¸õ

The system first break the given input into different syllables. So the input word become 3 syllables

Å - ½ì - ¸õ

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

3.5.Spectral Smoothening

In both speech synthesis and audio coding, there are circumstances where subsequent data segments have audibly different spectra at their adjoining boundaries. Signal processing can be used to smooth the existing waveform or create new data to bridge the gap between segments resulting from compression or coding errors. Straightforward linear interpolation in the frequency domain does not yield acceptable results, and therefore alternative algorithms are needed to provide more natural transitions. It is noted that spectral smoothing generally indicates modification of existing audio frames and spectral interpolation means the addition of frames; here we emphasize the addition of frames but do not distinguish between the two terms.

3.5.1.Optimal Coupling Technique

In concatenative synthesis that the boundaries of speech segments are fixed, but the optimal coupling technique allows the boundaries to move to provide the best fit with adjacent segments. A measure of mismatch is tested at a number of possible segment boundaries until the closest match is found.

While any form of measure may be used, for the sake of improving spectral quality, using a spectral discontinuity measure is appropriate. Measures considered include mel-frequency cepstral coefficient and the auditory-neural based measure. Fig. 3 shows an example scenario where moving the segment boundaries will noticeably change the spectral alignment of formants. In simple frame mismatch, distance measures are calculated for frames ending at various possible segment boundaries.

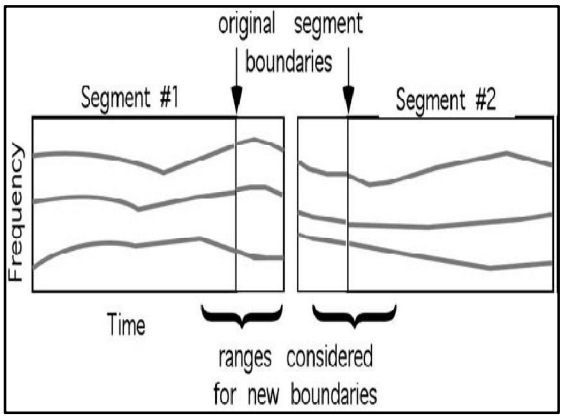

Figure 3: Optimal segment coupling

The distance measures take into account only the single audio frame from each speech segment which lies next to the boundary under consideration. Fig. 4 gives the Smoothened speech waveform for Tamil word ¦ºø¸¢È¡ý.

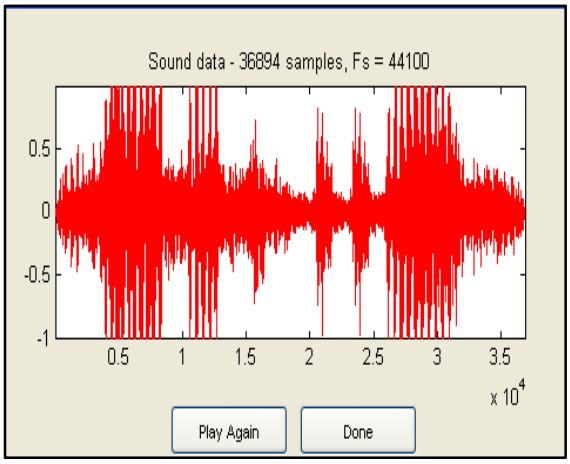

Figure 4: Smoothened speech waveform

coupling technique. The algorithm is conceptually simple and easy to implement. It can be combined with other spectral smoothing techniques and need not stand alone. Optimal coupling can successfully complement other smoothing techniques because it causes formants to be naturally closer to each other at segment boundaries. Since coupling does not modify the existing speech, it does not introduce additional artifacts. For spectral matching purposes, it effectively expands the number of speech segments in the database.

4.Results And Analysis

The execution of Tamil TTS System is explained in detail with the help of the interface designed for the system. The interface is designed with minimum complexity for the user in giving the text input for generation of speech. The front end is created using Matlab GUI report format. The interface allows us to enter the input as Tamil text through the keyboard or we can select the keys with mouse. The speech database is created from some speakers. The utterance of these sentences are collected in a noise free environment and are segmented into words and again segmented at syllables and are labeled automatically. The experiment is conducted by developing a speech database which includes all conversational sentences. It consist of 560 sentences of 2400 words and 1050 syllables. The speech samples are collected and syllabified and then labeled. Tests were conducted on general Tamil text and the synthesized speech is very close to natural speech.

Fig. 5 shows the GUI format for Dravidian - Tamil TTS with some modules like segmentation, speech concatenation, silence removal and spectral smoothening. The input text is converted into readable text and the syllabification module will generate the sequence of syllables that should be extracted from the speech corpus to be concatenated and play the sound files. Spectral smoothening with optimal coupling is to produce the continuous speech that resembles the natural human voice. The result is analyzed by using MOS (Mean Score Opinion)

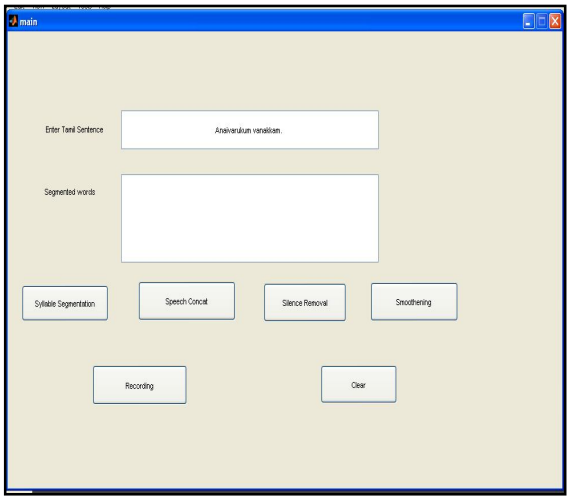

Figure 5: GUI format for Dravidian – Tamil TTS

4.1.Mean Opinion Score (MOS)

The Mean Opinion Score (MOS) provides a numerical indication of the perceived quality of received media after compression and/or transmission. The MOS is expressed as a single number in the range 1 to 5, where 1 is lowest perceived audio quality, and 5 is the highest perceived audio quality measurement. Five scores of subjective quality tests are 1-Bad, 2-Poor, 3- Fair, 4-Good, 5-Excellent.The MOS is generated by averaging the results of a set of standard, subjective tests where a number of listeners rate the heard audio quality of test sentences read aloud by both male and female speakers over the communications medium being tested. We took objective measurements from listener tests. Sixty expert listeners asked to indicate their preferences in terms of naturalness and intelligibility for different words and phrases. Table 1 shows the preliminary results of Mean Opinion Score (MOS).

Listeners	**MOS**
Male(28)	4.00
Female(32)	3.50

Table 1: Subjective test results

The listeners ranked the smoothened speech produced after concatenation as compared with natural speech. The results taken using these parameters signify that the speech is generated with minimal distortion.

5.CONCLUSION

Dravidian – Tamil TTS system using syllables as basic unit of concatenation is presented. The quality of the Synthesized speech is reasonably natural. The concatenative speech synthesis algorithm is an effective and easy method of synthesizing speech. In concatenative synthesizer, the designer provides recordings for phrases and individual words. The engine pastes the recordings together to speak out a sentence. The advantage of this mechanism is that it retains the naturalness and understandability of the speech. In silence removal part is to minimize the silence present in the speech signal. On the other hand the downside of this process is that it does not address the problem of spectral discontinuity. This problem can be reduced by the smoothening with optimal coupling technique. Smoothening is applied for smooth the transition between two speech segments

in order to produce the continuous speech that resembles the natural human voice. optimal coupling technique allows the boundaries to move to provide the best fit with adjacent segments. Basically in this technique a measure of mismatch is tested at a number of possible segment boundaries until the closest fit is found. Finally, We have observed that the efficiency and the performance of Tamil TTS is good natural speech.

6.Reference

1. 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.
2. Chopra D. , "Gayatri – A Fast Hindi Text To Speech System with Input Support For English Language", International Journal of Information Technology and Knowledge Management January-June 2011, Volume 4, No. 1, pp. 139- 141.
3. Veera Raghavendra, K. Prahallad, "A multilingual screen reader in Indian languages," In National Conference on Communications (NCC), Chennai, India, 2010.
4. Sreekanth Majji, Ramakrisshnan A.G " Festival Based maiden TTS system for Tamil Language", 3rd Language & TechnologyConference:Human Language Technologies as a Challenge for Computer Science and Linguistics October 5-7, 2007.
5. S. Schötz, "Data-driven formant synthesis of speaker age," In G. Ambrazaitis and S. Schötz (eds.). Lund Work-ing Papers 52, Proceedings of Fonetik, Lund, pp. 105–108. 2006.
6. Ganapathiraju M., Balakrishnan M., Balakrishnan N., Reddy R., "Om: One tool for many (Indian) languages," Journal of Zhejiang University Science, vol. 6A, no. 11, pp. 1348–1353, 2005.
7. Prahallad L., Prahallad K., Ganapathiraju M., "A simple approach for building transliteration editors for Indian languages," Journal of Zhejiang University Science, vol. 6A, no. 11, pp. 1354–1361, 2005.
8. Ramakrishnan, A.G. et.al., "Tools for the Development of a Hindi Speech Synthesis System", In 5th ISCA Speech Synthesis Workshop, Pittsburgh, pp. 109-114, 2004.
9. S. P. Kishore, R. Kumar, and R. Sangal "A data- driven synthesis approach for Indian languages using syllable as basic unit," In Intl. Conf. on Natural Language Processing (ICON), pp. 311–316, 2002.
10. Marian Macchi, Bellcore. Issues in text-to- speech synthesis. In Proc. IEEE International Joint Symposia on Intelligence and Systems, pp.318-325, 1998.