ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Developing Intonation Pattern for Tamil Text To Speech Synthesis System

K.C.Rajeswari, G. Karthick Prabu

Assistant Professor, Department of CSE, Sona College of Technology, Salem, TamilNadu, India[1]

M.E, Department of CSE, Sona College of Technology, Salem, TamilNadu, India[2]

ABSTRACT: The demands of interactive approaches to TTS require more freedom to express prosody than current systems allow. Most current TTS systems, including the Bell Labs TTS system, were designed to operate on text with little or no ‘‘mark-up’’ information beyond the text. The prosody subsystem was therefore designed conservatively. The application may be ‘‘intending’’ to convey that a set of words is a single proper noun, that a word is especially important, or that a word needs confirmation. This state information needs to be expressed prosodically, so one should think of speech synthesis more in the context of a concept-to-speech system than a text to- speech system. Similarly, there are applications where the simulation of emotions, subtle meanings in speech acts, and stylistic variations are desirable.

KEYWORDS: Simulation of emotions, prosody subsystem, Speech synthesis

I. Introduction

In the XXI century widespread use of computers opened a new stage in information interchange between the user and the computer. Among other things, an opportunity to input the information to computer through speech, and to reproduce in voice text information stored in the computer have been made possible. The paper is dedicated to the second part of this issue, i.e. to the computer-aided text-to-speech synthesis, that is recognized as a very urgent problem. Nowadays the solution of this problem can be applied in various fields. First of all, it would be of great importance for people with weak eyesight. In the modern world, it is practically impossible to live without an information exchange. The people with weak eyesight face with big problems while receiving the information through reading. A lot of methods are used to solve this problem. For example, the sound version of some books is created. As a result, people with weak eyesight have an opportunity to receive the information by listening. But there can be a case when the sound version of the necessary book couldn't be found. Therefore, the implementation of the speech technologies for information exchange for users with weak eyesight is of a crucial necessity. Synthesis of speech is the transformation of the text to speech. This transformation is converting the text to the synthetic speech that is as close to real speech as possible in compliance with the pronunciation norms of special language. TTS is intended to read electronic texts in the form of a book, and also to vocalize texts with the use of speech synthesis. When developing our system not only widely known modern methods but also a new approach of processing speech signal was used. In general, synthesis of speech can be necessary in all the cases when the addressee of the information is a person.

II. RELATED WORK

There have been three generations of speech synthesis systems. During the first generation (1962-1977) formant synthesis of phonemes was the dominant technology. This technology made use of the rules based on phonetic decomposition of sentence to formant frequency contours. The intelligibility and naturalness were poor in such synthesis. In the second generation of speech synthesis methods (from 1977 to 1992) the diphones were represented with the LPC parameters. It was shown that good intelligibility of synthetic speech could be reliably obtained from text input by concatenating the appropriate diphone units. The intelligibility improved over formant synthesis, but the naturalness of the synthetic speech remained low. The third generation of speech synthesis technology is the period from 1992 to the present day. This generation is marked by the method of "unit selection synthesis" which was introduced and perfected, by Sagisaka at ATR Labs. in Kyoto. The resulting synthetic speech of this period was close to humangenerated speech in terms of intelligibility and naturalness. Modern speech synthesis technologies involve quite complicated and sophisticated methods and algorithms. "Infovox" [2] speech synthesizer family is perhaps one of the best known multilingual text-to-speech products available today. The first commercial version, Infovox SA-101, was developed in Sweden at the Royal Institute of Technology in 1982 and it is based on formant synthesis. The latest full commercial version, Infovox 230, is available for American and British English, Chinese[6]. Digital Equipment Corporation [3] (DEC) talk system is originally descended from MITalk and Klattalk. The present version of the system is available for American English, German and Spanish and offers nine different voice personalities, four male, four female and one child. The present DECtalk system is based on digital formant synthesis. AT&T Bell Laboratories [4] (Lucent Technologies) also has very long traditions with speech synthesis. The first full textto-speech (TTS) system was demonstrated in Boston 1972 and released in 1973. It was based on articulatory model developed by Cecil Coker (Klatt 1987). The development process of the present concatenative synthesis system was started by Joseph Olive in mid 1970's (Bell Labs 1997)[15]. The current system is available for English, French, Spanish, Italian, German, Russian, Romanian, Chinese, and Japanese (Mcbius et al. 1996).

III.EXISTING SYSTEM

Text to speech synthesis is converting the text to the synthetic speech that is as close to real speech as possible according to the pronunciation norms of special language. Such systems are called text to speech (TTS) systems.

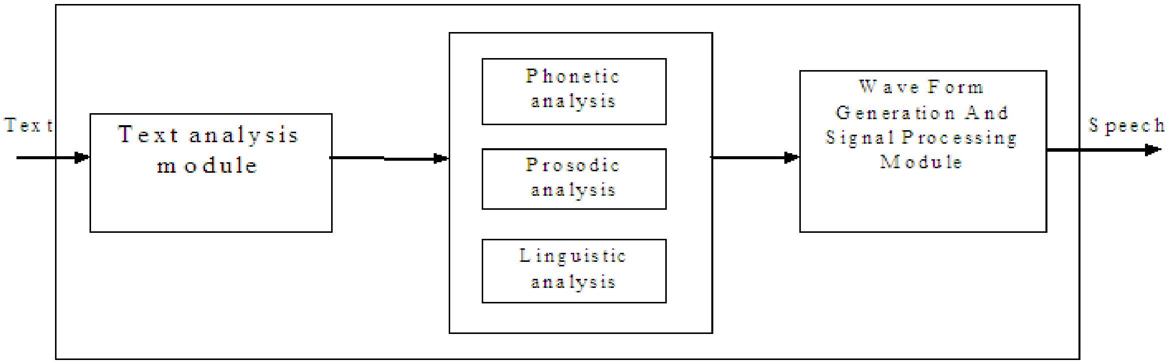

Fig.1 Block diagram for existing system

Input element of TTS system is a text, output element is synthetic speech. There are two possible cases. When it is necessary to pronounce the limited number of phrases (and their pronouncing linearly does not vary), the necessary speech material is simply recorded in advance. In this case, certain problems are originated. For example in this approach, it is not possible to sound the text, which is not known in advance. For this purpose the pronounced text has to be kept in computer memory. And it will lead to increase of the size of memory required for information content. This will bring to essential load of computer memory in case of much information and can create certain problems in operation. The main approach used in this paper is voicing of previously unknown text based on a specific algorithm. It is necessary to note that the approach to solving problem of speech synthesis essentially depends on the language for which it will be used and that the majority of currently available synthesizers basically were generated for English ,Spanish and Russian languages[16], and these synthesizers had not been applied to the Tamil language yet. Tamil language like as other Indian Languages(Bengali, Hindi) is an syllabic language[8]. In the view of the specificity of the Tamil language, the special approach is required. Every language has its own unique features. For example: there are certain contradictions between letters and certain sounds in English language. Thus, two different letters coming together, sound differently than when they are used separately. For example: letters (t), (h) separately do not sound the same as in chain (th). This is only one of problems faced in English language. In other words, the place of the letters affect on how they should be or should not be pronounced. Thus, according to the phonetic rules of English language the first letter (k) of the word (know) is not pronounced. As well as, Tamil language has certain pronunciation features. First of all, it should be noted that the letter (o) does not always pronounced like sound (o). There are some features based on phonetic rules of English language. For example: the first letter (o) in the word (out) is pronounced like (ah). From the foresaid it is clear that it the synthesizer programs developed especially for one language cannot be used in a different language, because the specificity of one language is not presumably typical for the others. Each program is based on algorithms corresponding to the phonetic rules of certain language. Till now there are no any programs of a synthesizer type that take into consideration the specificity of Tamil language. Two parameters, naturalness of sounding and intelligibility of speech, are applied for the assessment of the quality of synthesis system. One can say that naturalness of sounding of a speech synthesizer depends on how many generated sounds are close to natural human speech. By a intelligibility (ease for understanding) of a speech synthesizer is meant the easiness of artificial speech understanding. The ideal speech synthesizer should possess both characteristics: naturalness of sounding and intelligibility. Existing and being developed systems for speech synthesis are aimed at improvement of these two characteristics. The idea of combination of concatenation methods and of formant synthesis is the fundament of the system we developed. The rough, primary basis of a formed acoustic signal is created on the basis of concatenation of the fragments of an acoustic signal taken from speech of the speaker, i.e., a "donor". Then, this acoustic database is changed by the rules. The purpose of these rules is to give the necessary prosodies characteristics (frequency of the basic tone, duration and energy) to the "stuck together" fragments of an acoustic signal. The method of concatenation together with an adequate set of base elements of compilation provides for qualitative reproduction of spectral characteristics of a speech signal, and the set of rules provides for the possibility of generating natural intonation-prosodies mode of pronouncements. Formant synthesis does not use any samples of human speech. On the contrary, the speech message of the synthesized speech is created by means of acoustic model. Such parameters as own frequency, sounding and noise levels vary in order to generate natural form of a signal of artificial speech. In systems of prosody modeling (earlier it was called compilation), synthesis is carried out by sticking together necessary units from available acoustic units. Concatenation of segments of written speech lays in the basis of prosody modeling[1]. As a rule, prosody modeling gives naturalness to sounding of the synthesized speech. Nevertheless, the natural fluctuations in //speeches and the automated technologies of segmentation of speech signals create noise in the generated fragment and this decreases the naturalness of sounding. //An acoustic signal database (ASD)[19], which consists of fragments of a real acoustic signal, i.e. the elements of concatenation (EC), is the basis of any system of synthesis of the speech based on concatenation method. Dimension of these elements can be various depending on a concrete way of synthesis of speech; it can be phonemes, allophones, syllables, diaphones, words etc. In the system developed by us the elements of concatenation are diaphones and various combinations of vowels. But it is necessary to note that the study of generation of one-syllabic words consisting of four letters (stol, dörd) is still underway and that is why this words are included into the base as indivisible units. The speech units used in creation ASD are saved in WAV format. The process of creation of ASD consists of the following stages: Stage 1: In the initial stage, the speech database is created on the basis of the main speech units of the donor speaker. Stage 2: The speech units from speaker's speech are processed before being added into database. It's done in the following steps: a) Speech signals were sampled at 16 kHz and it makes possible to define the period T with a precision of 10-4. b) Removal of surrounding noise from the recorded speech units. For this purpose we use the algorithm of division of a phrase realization into speech and pauses. It is supposed, that the first 10 frames do not contain a speech signal. For this part of signal we calculate mean value and dispersion of Et and Zt and obtain statistical characteristics of noise. Stage 3: As described above, the ASD plays the main role in speech synthesis. The information stored in AED is used in different modules of synthesis. In our system, CE is stored in .wav format, with 16 kHz frequency. Each wav file includes the following elements:

0. The description of CE
1. The count of speech signal parts – N
2. Energy of speech signal – E
3. Amplitude of CE - A
4. The frequency of crossing zero– Z

Stage 4: At the following stage another corresponding variants of each CE are created. In spite of the fact that it increases the quantity of ASD elements, but at the same time it makes possible to reduce the quantity of modules for generation of a target signal.

The block of linguistic processing and the voicing module. At first, the input text is processed in the Linguistic block and the obtained phonemic transcript or is passed to the second block, i.e., to the Voicing block of system. In the Voicing block after certain stages the obtained speech signal is sounded.

IV. PROPOSED SYSTEM

4.1 Block of linguistic processing

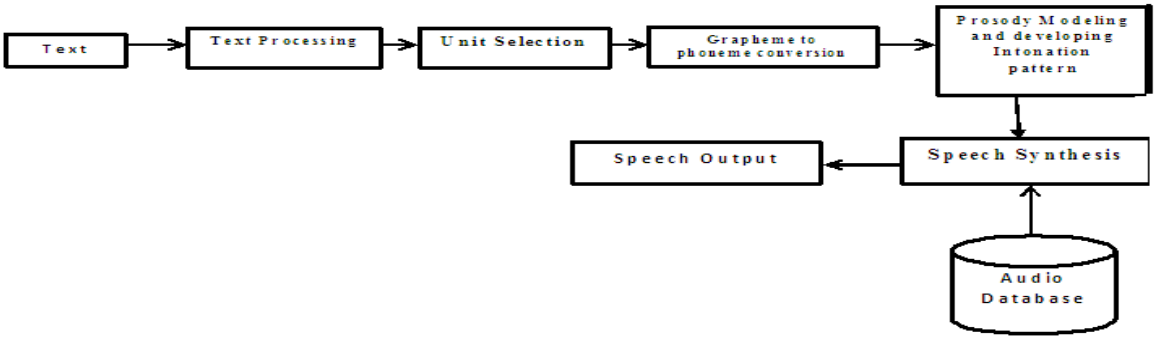
Text
Text Processing
Unit Selection
Grapheme to phoneme conversion
Prosody Modeling and developing Intonation pattern
Speech Synthesis
Speech Output
Audio Database

Fig.2 Block Diagram of Proposed System

4.1.1 Text input

The sounded text can be entered in any form. The size or font type is of no importance. The main requirement is that the text must be in Tamil language.

4.1.2 Initial text processing

For forming of transcriptional record, the input text should be shown as sequence of accentuated spelling words separated by space and allowed punctuation marks. Such text can conditionally be named as "normalized". Text normalization is a very important issue in TTS systems. The general structure of normalizer is explained in Figure 4. This module has several stages as it is shown in the figure.

Stage 1: Spell-checking of the text The spell-checkers are used in some cases (modules of correction of spelling and punctuation errors). The module helps to correct spelling errors in the text thereby to avoid voicing of these errors.

Stage 2: A pre-processing module A pre-processing module organizes the input sentences into manageable lists of words. First, text normalization isolates words in the text. For the most part this is as trivial as looking for a sequence of alphabetic characters, allowing for an occasional apostrophe and hyphen.

It identifies numbers, abbreviations, acronyms, and transforms them into full text when needed.

Stage 3: Number Expansion Text normalization then searches for numbers, times, dates, and other symbolic representations .These are analyzed and converted to words. Someone needs to code up the rules for the conversion of these symbols into words, since they differ depending upon the language and context.

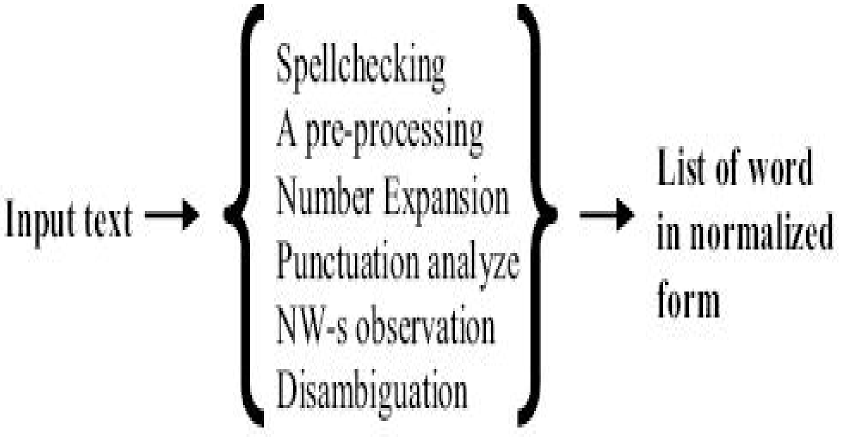

Fig.3 Text Normalization System

Stage 4: Punctuation analyze

Whatever remains is punctuation. The normalizer will have rules dictating if the punctuation causes a word to be spoken or if it is silent. (Example: Periods at the end of sentences are not spoken, but a period in an Internet address is spoken as "dot.") In normal writing, sentence boundaries are often signaled by terminal punctuation from the set: full stop, exclamation mark, question mark or comma {. ! ? ,} followed by white spaces. In reading a long sentence, speakers will normally break up the sentence into several phrases, each of which can be said to stand alone as an intonation unit. If punctuation is used liberally so that there are relatively few words between the commas, semicolons or periods, then a reasonable guess at an appropriate phrasing would be simply to break the sentence at the punctuation marks though this is not always appropriate. Hence, determining the sentence break and naming the type of sentence has to be done so as to apply the prosodic rules. In natural speech, speakers normally and naturally give
pauses between sentences

4.1.3 Unit Selection:

The unit selection is the basic steps for synthesis. They can be

Low concatenation distortion

Low prosodic distortion

General nature, if it's not restricted then text-to-speech is used

On the basis of the issue discussed before[9] , we are in need of considering the syllabus that has to be used as basic units. There are many possibilities which include phonemes, diphones or triphones. The possibilities of the unit are V, CV, VC, VCV, VCCV and VCCCV, where V stands for a vowel and C stands for a consonant. During concatenation the duration rules are applied which leads the database to contain only long vowels. The total number of units is around 2500.

4.1.4 Grapheme to phoneme conversion

The Grapheme to phoneme conversion in which the form of the word is given as input and thus the letter is converted into sound by using sound rules (That is the Phonetics)

4.1.5 Prosody Modelling:

The speech database is developed for prosody modeling.

The work that can be carried in the prosody modeling include following parameters

- Pitch
- Duration
- Intonation of the speech

The quality of the speech is improved by varying parameters the quality refers to the naturalness of the speech and pause that is given by the each word is developed[18].

4.1.6 Developing Intonation Pattern

This project proposed an intonation model using feed forward neural network (FFNN)[13] for syllable based text to speech (TTS) synthesis system for Tamil. The features used to model the neural network include set of positional, contextual and phonological features. The proposed intonation model predicts three F0 values correspond to initial, middle and final positions of each syllable. These three F0 values captures the broad shape of the F0 contour of the syllable.

4.1.7 Speech Synthesis:

In the speech synthesis[12] the recorded speech matches with the sound rules and they are check with given text and finally the output is given .

V. Conclusion and future work

On the above mentioned grounds, the voicing of words of any text in Tamil language is carried out with the help of a limited database set. In this study the framework of a TTS system for Tamil language is built. Although the system uses simple techniques it provides promising results for Tamil language, since the selected approach, namely the concatenative method, is very well suited for Tamil language. The system can be improved by improving the quality of the speech files recorded. In particular, the work on intonation is not finished because segmentation was made manually and there is noticeable noise in voicing. It is planned to apply independent segmentation and to improve the quality of synthesis in the future. The punctuations are removed in the preprocessing step just to eliminate some inconsistencies and obtain the core system. In the future versions of the TTS, the text can be synthesized in accordance with the punctuations for considering the emotions and intonations as partially achieved in some of the researches. The synthesis of a sentence ending with a question mark can have an interrogative intonation and synthesis of a sentence ending with an exclamation mark can be an amazing intonation. In addition to these, other punctuations can be helpful for approximating the synthesized speech to its human speech form such as pausing at the end of the sentences ending with full stop and also pausing after the punctuation comma. Major issues considered in developing TTS are text corpus collection, recording and labeling the speech corpus, deriving letter to sound rules and prosody modeling for Tamil language can be solved and attempts to produce a naturalness in speech. The dictionary-based approach is quick and accurate. The consistent evaluation of speech synthesis systems may be difficult because of a lack of universally agreed objective evaluation criteria. Different organizations often use different speech data. Recently, however, some researchers have started to evaluate speech synthesis systems using a common speech dataset.

References

1. Rajeswari K.C, and Uma Maheswari "Prosody Modeling Techniques for Text to Speech synthesis systems-A survey,international journal of computer applications,2012,vol.39 ,No.16
2. A.W. Black and K. Lenzo, "Building voices in the Festival speechsynthesis system", http://www.infovox.org, December 2009.
3. http://www.digital.com/
4. http://www.bell-labs.com/project/tts

5. Anton Batliner,et all "Prosody Models, Automatic speech understanding and speech synthesis: Towards the common ground", Proceedings of IEEE Inform 2001
6. Zhiyong WU et all, "Modeling prosody patterns for Chinese Expressive Text-to Speech Synthesis", IEEE Trans on Software Engineering,2012.
7. Chung-Hseien Wu, et all, "Hierarchical Prosody Conversion Using Regression-Based Clustering for Emotional Speech Synthesis",Proceedings of the 10th IEEE International Conference on Computing and Communications, Dalian,China, 2010, pp.5-13.
8. Vinodh.M.V,et all, "Using polysyllabic units for Text to Speech Synthesis in Indian Languages", Computer Engineering, 2010, vol. 31, No. 6.
9. K.Parthasarathy,et all, "A research Bed for Unit Selection Based Text To Speech Synthesis" , IEEE, Aug 2008.
10. Jing Zhu , et all, "Intonation and Prosody Conversion for Expressive Mandarin Speech Synthesis", ICSP/IEEE 2012, PP. 268-273.
11. M.Szymanski, et all, "Optimization of Unit Selection Speech Synthesis", Proceedings of ICPhs Hongkong 2011.
12. Yousef Tabet, "Speech Synthesis Techniques A Survey" in Proc. ICSLP, 2012.
13. V.Ramu Reddy, et all, "Two-stage intonation modeling using feed forward neural networks for syllable based text-to-speech synthesis", Elsevier, N.4, Feb 2013.
14. Abhijeet Sanwan,et all "Automatic language analysis and identification based on speech production knowledge", Proceedings of 20th IEEE International Conference 2010.
15. Ashwin Bellur,et all"Prosody Modeling for Syllable-Based Concatenative Speech Synthesis of Hindi and Tamil", Proceedings of IEEE Inform,2011
16. Heiga Zen,et all, "The HMM-based Speech Synthesis System (HTS) Version 2.0", IEEE Trans on Software Engineering,2011.
17. M.Szyman,et all "Optimization Of Unit Selection Speech Synthesis,", IEEE Trans on Software Engineering,2011. Nov, 2011.
18. Dan-ning Jiang,et all "Prosody Analysis and Modeling for Emotional Speech Synthesis", Proceedings of 20th IEEE International Conference, 2011.
19. Norbert Braunschweiler, "The Prosodizer- Automatic Prosodic Annotations of speech Synthesis Database", Proceedings of 16th IEEE International Conference Computer Associates, 2006, PP.1-14.