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. 3, Issue 9, September 2015

Implementation of Text to Speech Conversion Technique

Mohd Bilal Ganai[1], Er jyoti Arora[2]

M.Tech Student, Dept. of Computer Science, Desh Bhagat University, Punjab, India[1]

Assistant Professor, Dept. of Computer Science, Desh Bhagat University, Punjab, India [2]

ABSTRACT: Text-to-speech (TTS) is the generation of synthesized speech from text. Our goal is to make synthesized speech as intelligible, natural and pleasant to listen, as human speech. Speech is the primary means of communication between people. During synthesis very small segments of recorded human speech are concatenated together to produce the synthesized speech. The quality of a speech synthesizer is judged by its similarity to the human voice and by its ability to be understood. A text-to-speech synthesizer allows people with visual impairments and reading disabilities to listen to written works on a home computer. Many computer operating systems have included speech synthesizers since the early 1990s. Recent progress in speech synthesis has produced synthesizers with very high intelligibility but the sound quality and naturalness still remain a major problem. However, the quality of present products has reached an adequate level for several applications, such as multimedia and telecommunications. The following thesis presents a brief overview of the main text-to-speech synthesis problems, and the initial work done in building a TTS in English.

KEYWORDS: Text-to speech conversion, phoneme, Synthesis, concatenation.

I. INTRODUCTION

Language is the ability to express one’s thoughts by means of a set of signs (text), gestures, and sounds. It is a distinctive feature of human beings, who are the only creatures to use such a system. Speech is the oldest means of communication between people and it is also the most widely used ‘Speech synthesis’ also called ‘Text to speech synthesis’ is the artificial production of human speech. A computer system used for this purpose is called a speech synthesizer and can be implemented in software. A text-to-speech (TTS) system converts text to speech. At first sight, this task does not look too hard to perform. After all we all have a deep knowledge of reading rules of our mother tongue. They were transmitted to us, in a simplified form, at primary school, and we improved them year after year. But in the context of TTS synthesis, it is impossible to record and store all the words of the language. Some other method has to be implemented for this purpose.

The quality of a speech synthesizer is judged by its similarity to the human voice and by its ability to be understood. A text-to-speech synthesizer allows people with visual impairments and reading disabilities to listen to written works on a home computer. Many computer operating systems have included speech synthesizers since the early 1990s. Astro-physician Stephen Hawkins, who is completely paralyzed, gives all his lectures using a TTS system.
This project gives an idea about developing a pc based text-to speech conversion technique using MATLAB.

SPEECH SYNTHESIS

1.1What is speech synthesis?

A Text-To-Speech (TTS) synthesizer is a computer-based system that should be able to read any text aloud, when it is directly introduced in the computer by an operator. It is more suitable to define Text-To-Speech or speech synthesis as an automatic production of Speech by ‘grapheme to phoneme’ transcription. A grapheme is the smallest distinguishing unit in a written language. It does not carry meaning by itself. Graphemes include alphabetic letters, numerical digits, punctuation marks, and the individual symbols of any of the world's writing systems. A phoneme is "the smallest segmental unit of sound employed to form meaningful utterances".

DOI: 10.15680/IJIRCCE.2015. 0309075

1.2 PHONETICS

In most languages the written text does not correspond to its pronunciation. So that in order to describe correct pronunciation some kind of symbolic presentation is needed. Every language has a different phonetic alphabet and a different set of possible phonemes and their combinations. The number of phonetic symbols is between 20 and 60 in each language. "A set of phonemes can be defined as the minimum number of symbols needed to describe every possible word in a language". In English there are about 44 phonemes. Due to complexity and different kind of definitions, the number of phonemes in English and most of the other languages cannot be defined exactly. Phonemes are abstract units and their pronunciation depends on contextual effects, speaker's characteristics, and emotions. During continuous speech, the articulator movements depend on the preceding and the following phonemes. The articulators are in different position depending on the preceding one and they are preparing to the following phoneme in advance. This causes some variations on how the individual phoneme is pronounced. These variations are called allophones which are the subset of phonemes and the effect is known as co -articulation. For example, a word lice contains a light /l/ and small contains a dark /l/. These l's are the same phoneme but different allophones and have different vocal tract configurations. Another reason why the phonetic representation is not perfect, is that speech signal is always continuous and phonetic notation is always discrete. The phonetic alphabet is usually divided in two main categories, vowels and consonants. Vowels are always voiced sounds and they are produced with the vocal cords in vibration, while consonants may be either voiced or unvoiced. Vowels have considerably higher amplitude than consonants and they are also more stable and easier to analyze and describe acoustically. Because consonants involve very rapid changes they are more difficult to synthesize properly.
Few examples of different phonetic notations are shown below.

1.3 SYNTHESIZER TECHNOLOGY

The most important qualities of a speech synthesis system are naturalness and intelligibility. Naturalness describes how closely the output sounds like human speech, while intelligibility is the ease with which the output is understood. The ideal speech synthesizer is both natural and intelligible. Speech synthesis systems usually try to maximize both characteristics.
The primary technology for generating synthetic speech is concatenative synthesis.

1.3.1 Concatenative synthesis

Concatenative synthesis is based on the concatenation (or stringing together) of segments of recorded speech. Connecting prerecorded natural utterances is probably the easiest way to produce intelligible and natural sounding synthetic speech. However, concatenative synthesizers are usually limited to one speaker and one voice and usually require more memory capacity.

One of the most important aspects in concatenative synthesis is to find correct unit length. The selection is usually a trade-off between longer and shorter units.
With longer units' high naturalness, less concatenation points are achieved, but the amount of required units and memory is increased. With shorter units, less memory is needed, but the sample collecting and labeling procedures become more difficult and complex. In present systems units used are usually words, syllables, phonemes.

1.3.2 Domain-specific synthesis

Word is perhaps the most natural unit for written text and some messaging systems with very limited vocabulary. Concatenation of words is relative easy to perform.

Domain-specific synthesis concatenates prerecorded words and phrases to create complete utterances. It is used in applications where the system's output is limited to a particular domain, like transit schedule announcements or weather reports. The technology is very simple to implement, and has been in commercial use for a long time, in devices like talking clocks and calculators. The level of naturalness of these systems can be very high because the variety of sentence types is limited, and they closely match the prosody and intonation of the original recordings.

Because these systems are limited by the words and phrases in their databases, they are not general- purpose and can only synthesize the combinations of words and phrases with which they have been preprogrammed. However, there is a great difference with words spoken in isolation and in continuous sentence which makes the continuous speech to sound very unnatural. Because there are hundreds of thousands of different words and proper names in each language, it is quite clear that we cannot create a database of all words and common names in the world and so word is not a suitable unit for any kind of unrestricted TTS system.

Thus, for unrestricted speech synthesis (text-to- speech) we have to use shorter pieces of speech signal, such as syllables, phonemes or even shorter segments.

Phonemes are probably the most commonly used units in speech synthesis because they are the normal linguistic presentation of speech. The inventory of basic units is usually between 40 and 50, which is clearly the smallest compared to other units. Using phonemes gives maximum flexibility.

1.4 WORD PRONUNCIATION

Character to voice is not a really big task. This is because there are only 26 characters in English and each character has a unique pronunciation. However when we have to read lengthy texts, character to voice is not recommended at the user level, as it is difficult to make out a word from the characters read.
As we have played the wave file corresponding to every character read, in character to voice conversion, we can also play the wave file for every word read. But practically it is impossible to record all the words of a dictionary. Hence there is a need to think of some other alternative. In the first attempt to play a word as a whole we can think of playing syllables of a word.

II. RELATED WORK

[1] In this paper 2014.Inoue, Takuma et.al. [1] has been worked on hybrid text-to-speech based on sub-band approach. This paper proposes a sub-band speech synthesis approach to develop high-quality Text-to-Speech (TTS). For the low-frequency band and high-frequency band, Hidden Markov Model (HMM)-based speech synthesis and waveform-based speech synthesis are used, respectively. Both speech synthesis methods are widely known to show good performance and to have benefits and shortcomings from different points of view. One motivation is to apply the right speech synthesis method in the right frequency band. Experiment results show that in terms of the smoothness the proposed approach shows better performance than waveform-based speech synthesis, and in terms of the clarity it shows better than HMM-based speech synthesis.

[2] 2013. Schultz, Tanjaet. al. [3] has been implemented the GlobalPhone. The global phone is a multilingual text & speech database in 20 languages. This paper describes the advances in the multilingual text and speech database Global Phone, a multilingual database of high quality read speech with corresponding transcriptions and pronunciation dictionaries in 20 languages. Global Phone was designed to be uniform across languages with respect to the amount of data, speech quality, the collection scenario, the transcription and phone set conventions. With more than 400 hours of transcribed audio data from more than 2000 native speakers Global Phone supplies an excellent basis for research in the areas of multilingual speech recognition, rapid deployment of speech processing systems to yet unsupported languages, language identification tasks, speaker recognition in multiple languages, multilingual speech synthesis, as well as monolingual speech recognition in a large variety of languages.

III. PROBLEMS IN SPEECH SYNTHESIS

The problem area in speech synthesis is very wide. There are several problems in text pre- processing, such as numerals, abbreviations, and acronyms
This chapter describes the major problems in text-to-speech research.

1) Text- to-Phonetic Conversion

The first task faced by any TTS system is the conversion of input text into linguistic representation, usually called text-to-phonetic or grapheme-to- phoneme conversion. The difficulty of conversion is highly language depended and includes many problems. In some languages, such as Hindi or Telugu, the conversion is quite simple because written text almost corresponds to its pronunciation. For English and most of the other languages the conversion is much more complicated. A very large set of different rules and their exceptions is needed to produce correct pronunciation and prosody for synthesized speech.

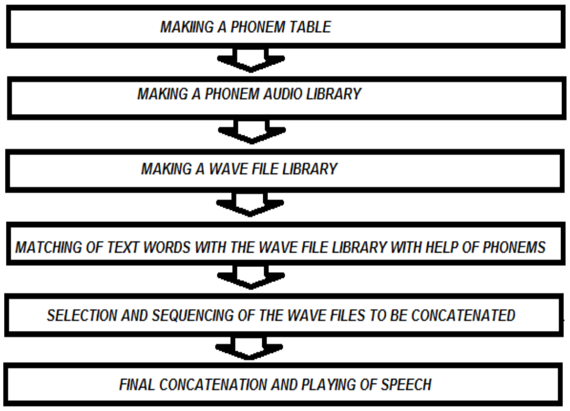
2) Pronunciation

The second task is to find correct pronunciation for different contexts in the text. Some words, called homographs, cause the most difficult problems in TTS systems. Homographs are spelled the same way but they differ in meaning and usually in pronunciation (e.g. fair, lives). The word lives is for example pronounced differently in sentences "Three lives were lost" and "One lives to eat". Some words, e.g. present, has different pronunciations depending on the context. (I was present there when he received the present).

The characters 'th' in ‘mother’ and ‘think’ is pronounced differently. Some sounds may also be either voiced or unvoiced in different context. For example, phoneme /s/ in word dogs is voiced, but unvoiced in word cats.
Finding correct pronunciation for proper names, especially when they are borrowed from other languages, is usually one of the most difficult tasks for any TTS system. Unfortunately, it is clear that there is no way to build a database of all proper names in the world.

IV. PROPOSED WORK

The literature would be studied in detail on the text to speech conversion technique, speech wave generation or concatenation techniques in order to know their workflow. The literature survey will be carefully conducted to find the research gaps in the existing speech waveform techniques. Then the proposed model will be designed and improved to remove the shortcomings and research gaps of the existing schemes. The proposed model will be then implemented using the MATLAB. The text-to-speech conversion system is coded in MATLAB. The system is developed by creating a library of phonemes, a library of phoneme audio files and a dictionary of words with their phoneme representation.


MAKIING A PHONEM TABLE
MAKING A PHONEM AUDIO LIBRARY
MAKING A WAVE FILE LIBRARY
MATCHING OF TEXT WORDS WITH THE WAVE FILE LIBRARY WITH HELP OF PHONEMS
SELECTION AND SEQUENCING OF THE WAVE FILES TO BE CONCATENATED
FINAL CONCATENATION AND PLAYING OF SPEECH

Flow Chart of proposed work

The system generates takes a sentence, analyze each word and find out their corresponding phonemes. Then it concatenates all the phonemes from the phoneme audio library and then plays the audio which sounds like a speech of sentence. It also displays a waveform and spectrum of the generated speech sound.

V. APPLICATIONS OF SYNTHETIC SPEECH

Synthetic speech may be used in several applications. some applications, such as reading machines for the blind or electronic-mail readers, require unlimited vocabulary and a TTS system is needed.

The application field of synthetic speech is expanding fast whilst the quality of TTS systems is also increasing steadily.

Speech synthesis systems are also becoming more affordable for common customers, which makes these systems more suitable for everyday use. For example, better availability of TTS systems may increase employing possibilities for people with communication difficulties.

Applications for the Blind

Probably the most important and useful application field in speech synthesis is the reading and communication aids for the blind. Before synthesized speech, specific audio books were used where the content of the book was read into audio tape. It is clear that making such spoken copy of any large book takes several months and is very expensive. It is also easier to get information from computer with speech instead of using special bliss symbol keyboard, which is an interface for reading the Braille characters.

A blind person cannot also see the length of an input text when starting to listen it with a speech synthesizer, so an important feature is to give in advance some information of the text to be read. For example, the synthesizer may check the document and calculate the estimated duration of reading and speak it to the listener. Also the information of bold or underlined text may be given by for example with slight change of intonation or loudness.

Applications for the Deafened and Vocally Handicapped

People who are born-deaf cannot learn to speak properly and people with hearing difficulties have usually speaking difficulties. Synthesized speech gives the deafened and vocally handicapped an opportunity to communicate with people who do not understand the sign language.

Educational Applications

Synthesized speech can be used also in many educational situations. A computer with speech synthesizer can teach 24 hours a day and 365 days a year. It can be programmed for special tasks like spelling and pronunciation teaching for different languages. It can also be used with interactive educational applications.

Applications for Telecommunications and Multimedia

The newest applications in speech synthesis are in the area of multimedia. Electronic mail has become very usual in last few years. However, it is sometimes impossible to read those E-mail messages when being for example abroad. There may be no proper computer available or some security problems exist. With synthetic speech e-mail messages may be listened to via normal telephone line. Synthesized speech may also be used to speak out short text messages (sms) in mobile phones.

Other Applications

In principle, speech synthesis may be used in all kind of human-machine interactions. For example, in warning and alarm systems synthesized speech may be used to give more accurate information of the current situation. Using speech instead of warning lights or buzzers gives an opportunity to reach the warning signal for example from a different room

VI. EXPERIMENTS AND RESULTS

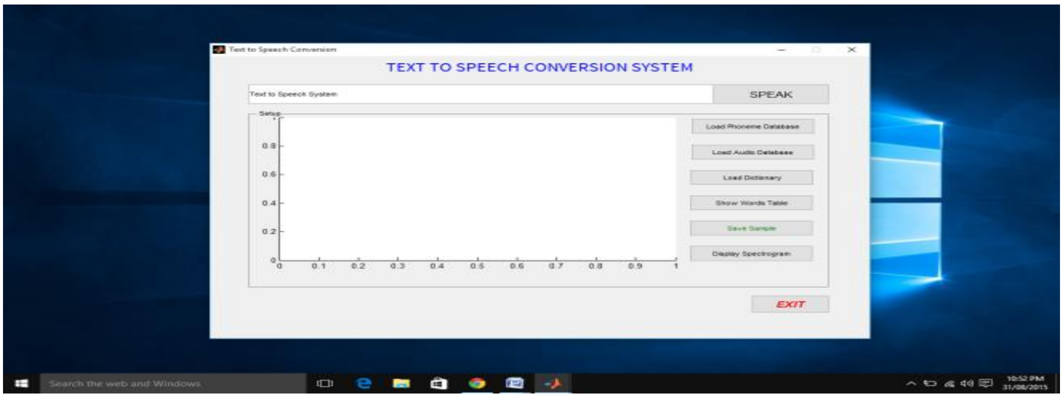
The text-to-speech conversion system is coded in MATLAB. The system is developed by creating a library of phonemes, a library of phoneme audio files and a dictionary of words with their phoneme representation. The system generates takes a sentence, analyze each word and find out their corresponding phonemes. Then it concatenates all the phonemes from the phoneme audio library and then plays the audio which sounds like a speech of sentence. It also displays a waveform and spectrum of the generated speech sound.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

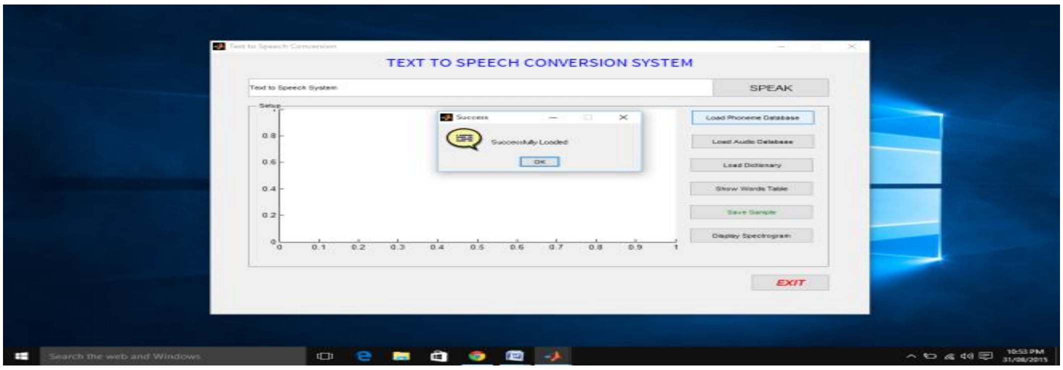
Vol. 3, Issue 9, September 2015



Text to Speech System

The above figure depicts all functions & their conversions with the given parameters. We can check with the statics of objects.

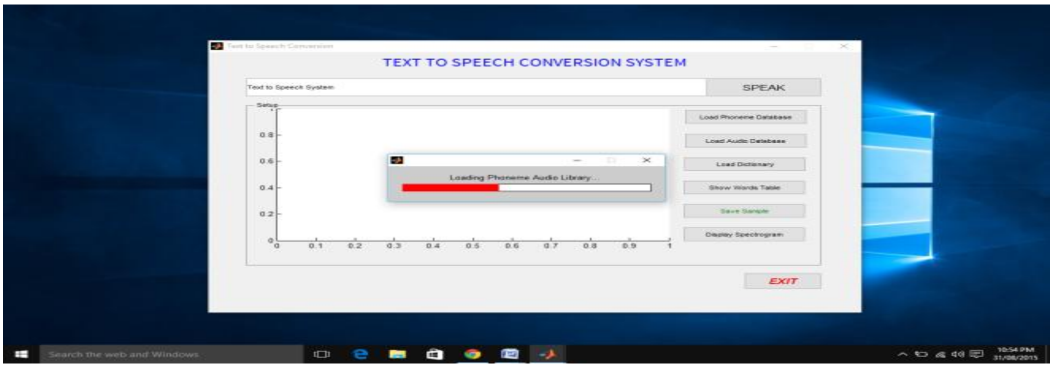
Figure 1: The Front Screen of the Text to Speech Conversion System Developed in MATLAB



Successfully Loaded

The above diagram depicts the creation of Phoneme Database.

Figure 2: Phoneme Table is Loaded Successfully

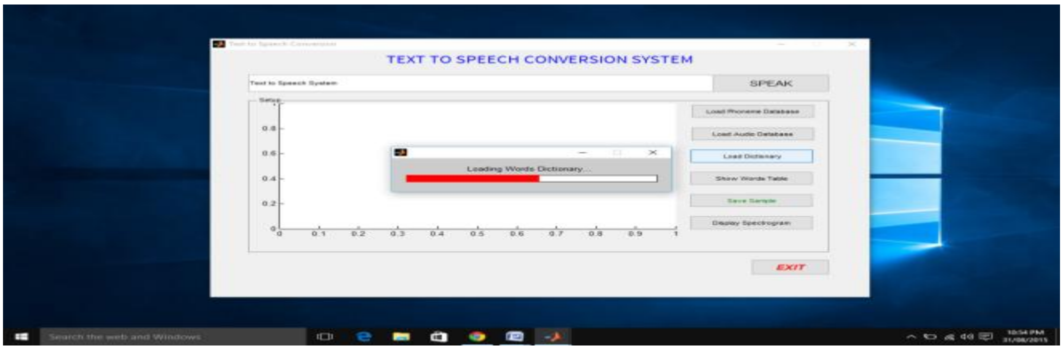


Loading Phoneme Audio Library...

The above diagram depicts the creation of audio database.

Figure 3: Phoneme Audio Library is loading

DOI: 10.15680/IJIRCCE.2015. 0309075

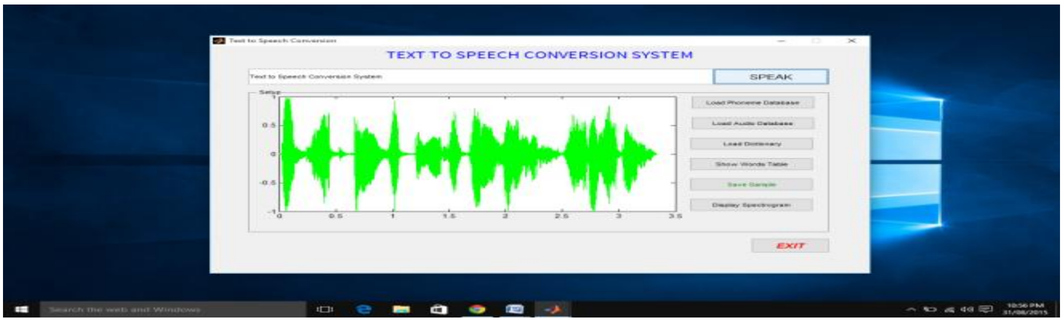


Text to Speech System

Loading Words Dictionary...

The above diagram shows word table database

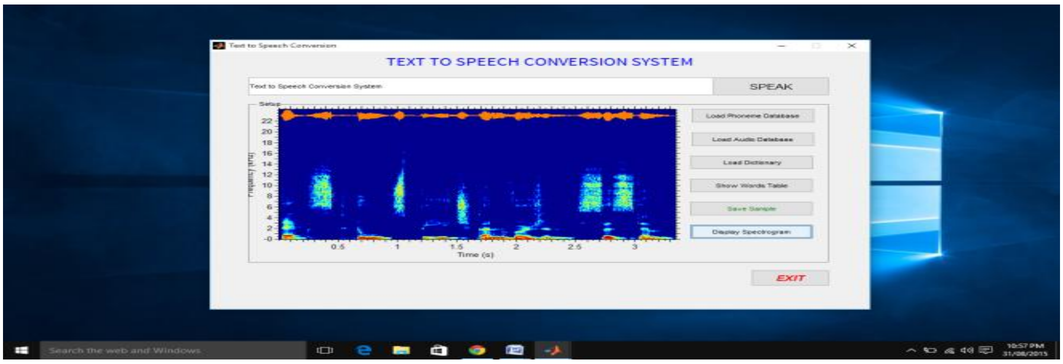
Figure 4: Words Library is loading



Text to Speech Conversion System

The above diagram is creating a waveform for given words of library.

Figure 5: The Audio Waveform of the Spoken Text: “Text to Speech Conversion System”



Text to Speech Conversion System

The color combination is showing the quality for the words in library.

Figure 6: Spectrogram of the Spoken Text

VII. CONCLUSION

As described in this paper, we have implemented a text to speech conversion system using concatenated phoneme library techniques. We developed a system which produces more human like speech using recorded phonemes. The system is an advantage over available text to speech conversion system in the way our system produces the sound which is very much human like. The system is analysed qualitatively.

FUTURE WORK

The future scope of the project is to add more emphasis simulation by making the speech sound with emotions. The flow of the text can be improved by adding more information about the context.

REFERENCES

1. Adiga, Nagaraj, and S. R. MahadevaPrasanna. "A hybrid Text-to-Speech synthesis using vowel and non vowel like regions." In India Conference (INDICON), 2014 Annual IEEE, pp. 1-5. IEEE, 2014.
2. www.ims.unistuttgart.de/~moehler/synthspeech/
3. Neumann, Lukáš, and JiříMatas. "A real-time scene text to speech system." In ComputerVisionECCV2012.Workshopsand Demonstrations, pp. 619-622. Springer Berlin Heidelberg, 2012.
4. www.acoustics.hut.fi/publications/files/thess /lemmetty_mst/contents.html
5. http://books.google.co.in/books/about/An_Introduction_to_Text_To_Speech_Synthe.htL
6. .http://www.abelard.org/briefings/phonetic_chart_br itish_english.php
7. Schultz, Tanja, Ngoc Thang Vu, and Tim Schlippe. "GlobalPhone: A multilingual text & speech database in 20 languages." In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pp. 8126-8130. IEEE, 2013.
8. Toda, K. Tokuda, "Speech parameter generation algorithm considering global variance for HMM-based speech synthesis," Proc. of Eurospeech'05, pp. 2801–2804, 2005.
9. Toth, J. ; Kondelova, A. ; Rozinaj, G.. "Natural language processing of abbreviations". 2011. ELMAR, 2011 Proceedings.
10. Cuiying Yan ; Kai GAO ; Mei Li. "Processing natural language based query and context sensitive spelling suggestion in information retrieval". 2013. Modelling, Identification & Control (ICMIC), 2013 Proceedings of International Conference.
11. Haris, S.S. ; Omar, N.. "A rule-based approach in Bloom's Taxonomy question classification through natural language processing".2012. Computing and Convergence Technology (ICCCT), 2012 7th International Conference.
12. Weischedel, R.M.. "Knowledge representation and natural language processing". 2005. Proceedings of the IEEE

DOI: 10.15680/IJIRCCE.2015. 0309075