International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE)
Volume 3, Issue 2, February 2014

High Quality Text to Speech Synthesizer using Phonetic Integration

Mrs. S. D. Suryawanshi, Mrs. R. R. Itkarkar, Mr. D. T. Mane

Abstract— **A Text-To-Speech (TTS) synthesizer is computer based system that should be able to read any text aloud, whether it was straight bring in the computer by an operator or scanned and submitted to an Optical Character Recognition system. In the context of TTS synthesis, it is very complicated to record and accumulate all the words of the language. So it is in effect more appropriate to define TTS as the automatic production of speech by using the concept of grapheme and phonemes text of the sentences to complete. In this paper, we are implemented natural prosody generation in TTS for English using the phonetics integration. Here we use a method to design a Text to Speech conversion module by the use of Matlab by simple matrix operations. Firstly by the use of microphone some similar sounding words are recorded which contain the all phonemes of English language and recorded sounds are saved in .wav format in the directory. The recorded word sounds are then extracted and the sampled values are taken and separated into their basic phonetics. The Extracted phoneme are use to compare with input sentence phonemes and then concatenated phonemes to reconstruct the desired words. This method is simple to implement and involves much lesser use of memory spaces**

Keywords- **Text to Speech, Syllabification, Phonetic Concatenation, Text normalization**

I. INTRODUCTION

Synthesized speech can be created by concatenating part of recorded speech which is stored in a database. The power of a speech synthesizer is moderator by its similarity to the human being voice, and by its ability to be understood. The mainly significant qualities of a speech synthesis system are naturalness and Intelligibility. Naturalness expresses how intimately the output sounds like human speech, whereas intelligibility is the easiness with which the output is understood. The perfect speech synthesizer is providing both natural and intelligible speech hence speech synthesis systems usually try to maximize both characteristics. There are different significant factors to be considered while designing a Text to speech system that will produce clear speech[1]. A high quality TTS system would enable automatic sound broadcasts such as stock market reports and even automatic TV broadcasts by combining computer graphic animations generated from a script concatenative speech synthesis system that can generate high quality synthesized speech output. Here we focus on two aspects of the concatenative speech synthesis system: Phoneme Extraction, and the Phonetic concatenation. Synthesizer is used in many purpose and helped users hugely, especially those who need special care and support (such as blind, deafened and vocally handicapped) also speech synthesis help in education and excessive need for computers [2,3].

The following section tells the database containing recorded similar sound for each phoneme, In phoneme Extraction based on the phonemes, respective phoneme are chosen. These phoneme sounds are then concatenated to generate the wav file, which has synthesized speech, describe the whole sentence.

II. TEXT TO SPEECH SYSTEM

TTS Synthesizer is a computer based system that should be understand any text clearly whether it was establish in the computer by an operator or scanned and submitted to an Optical Character Recognition (OCR) system. The intention of a text to speech system is to convert an random given wording into a speak waveform. Most important workings of text to speech system are Text processing and Speech production[4].

The two primary methods for producing synthetic speech waveforms are concatenative synthesis and formant synthesis. We are used Concatenative synthesis for our TTS. Concatenative synthesis is stand on the concatenation of piece of recorded words. Usually concatenative synthesis constructs the most normal sounding synthesized words.

III. SPEECH GENERATION COMPONENT

Given order of phonemes, the idea of the speech generation component is to synthesize the acoustic waveform Speech generation has been attempted by concatenating the recorded words parts [4]. Recent state of art language synthesis produces natural sounding speech by using huge amount of speech pieces. Storage of huge number of pieces and their retrieval in real time is feasible due to availability of cheap memory and computation power. The problem related to the unit selection speech synthesis system are consider in three things that are Choice of unit size, generation of speech database and criteria for selection of a unit.

IV. SPEECH SYNTHESIS PROCESS

This TTS system is able to read any written text. This procedure is called text normalization, preprocessing and tokenization. In this system, we have developed a phonetic based text to speech synthesis system. We can improve the speech quality using MATLAB language [5, 6]. Fig 1 shows the block diagram for TTs system and Fig 2 shows the flowchart for the same. The details are explained step by in following section.

ISSN: 2278 – 909X

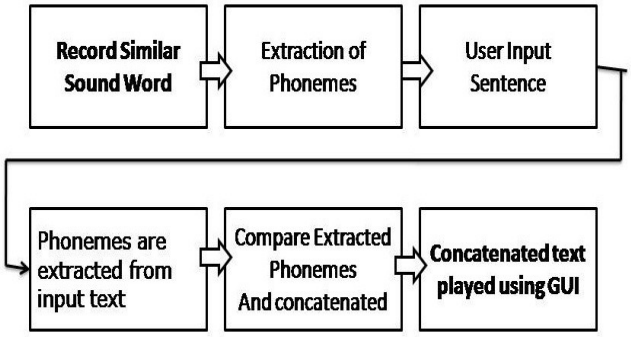

Fig1. Block Diagram for Text to speech Synthesis

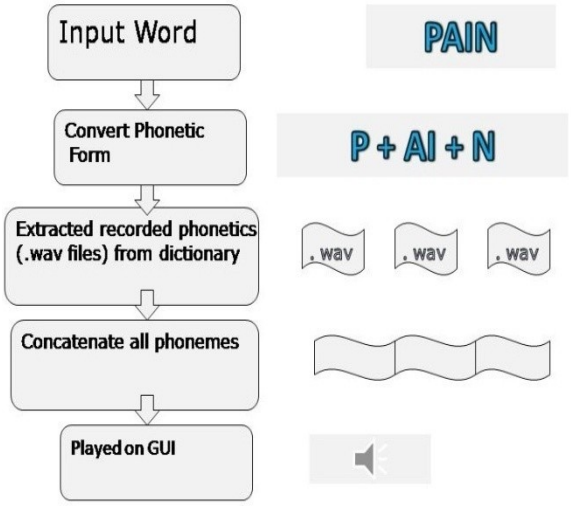

Fig 2. Flow chart for TTS with example

A. RECORDING SOUND USING MATLAB CODE
First using Microphone we save the Sound. After which that Sound is Stored in Matlab Directory with '.wav' extension. We can record the sound in Matlab using GUI i.e Graphical User Interface and the duration of the sound can also be adjusted. The sound which is recorded is displayed on the upper plot in Goethe spectrogram can be viewed when the sound is recorded and from that spectrogram we can detect the phonemes[5,6].

B. BRINGING THE SOUND INTO MATLAB WORK DIRECTORY
Using wavered function recorded sound are transferred into mat lab workspace Sample Code [Y, FS1, NS1] = wavread ('pain.wav');
Where:
Y = Variable used for storing the .wav file.
FS1 = Sampling frequency.
NS1 = Number of bits in each sample.

C. EXTRACTION OF PHONEMES
In English Language there are 44 phonemes, the selection of these 44 phonemes are based on the list from the source namely 'Orchestrating Success in Reading' by Dawn Reithaug under National Right to Read. The main aim of 44 phonemes is to extract the sound and from this we can create any English word [7]. Table 1 listing the some of English phonemes with examples.

Phoneme	Example
a	Cat ,rat, mat, sat
ae	Gain, Pain ,gate ,station
ee	Sweet ,heat, Meet ,these
ie	Tried, light ,my, might
ou	Road, blow, bone ,cold

(1)

Table 1. English phonemes with example

D. CONCATENATION
These phonemes can be concatenated to give various words. For example, suppose from word 'Gain' sound 'G' and 'ain' can be extracted and stored, when the word needed these two phonemes are concatenated and placed[5].

E. PHONEMES ARE EXTRACTED FROM INPUT TEXT
Whenever the user inputs a word, that word is found from the list, because of Phonetic Representation which is saved in another variable. Such as, phonetic representation of the input word 'pain' is 'p/ae/n' and if 'p' and 'ae' and ' n' are concatenated they give a sound of the word 'pain'. For removing the pause and frequency span between words we can use two filters i.e. 'Savitzky - Golay Smoothing filter ' and 'Median Filter ' whenever frequency span is large then it is used. Fig 3 shows these whose process for pain word. By using sliders we can adjust speech volume and frequency span of sentence which is shown in Fig 4.

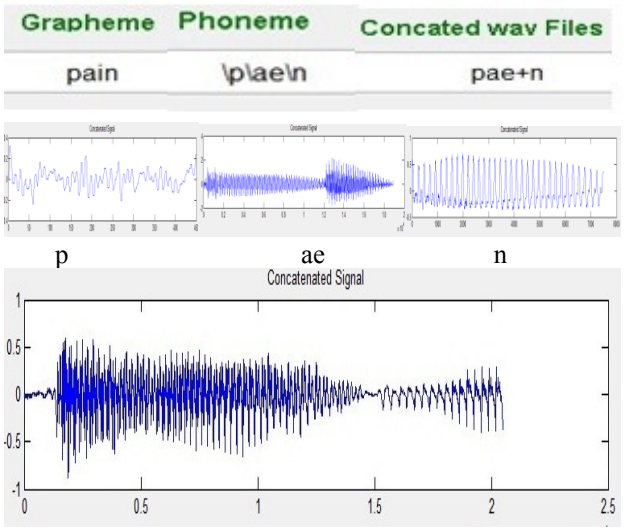

Fig.3. Spectrogram of pain word after concatenation

ISSN: 2278 – 909X

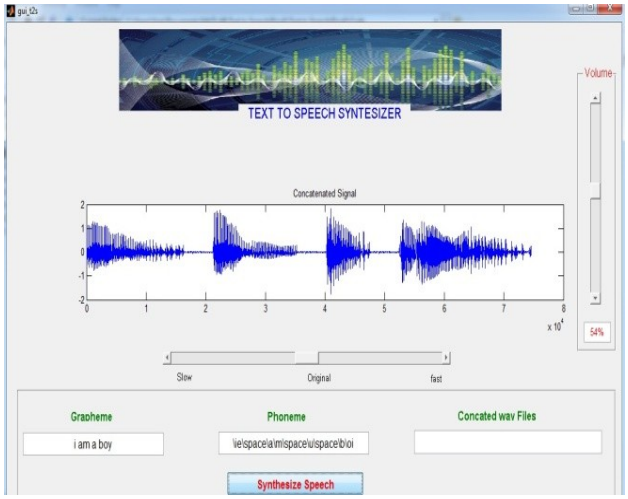

Fig.4. Spectrogram of sentence" I am a boy" after concatenation

V. QUALITY TEST

To evaluate the quality of the synthetic right to be heard produced by the developed system we passed out proper listening tests. Voice superiority testing is carried out using subjective test. In subjective tests, individual listeners listen to and rank the superiority of processed tone files according to a positive level [5]. Mean Opinion Score MOS is composed of five scores of subjective quality, 1-Bad, 2-Poor, 3- Fair, 4-Good, 5-Excellent. The MOS test was carried out by synthesizing a set of 50 sentences that were selected from the speech corpus randomly and did not participate in the training set. 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 [8], here tests are conducted with ten students with the different age group of 20 to 26 years. The tests were conducted in the laboratory environment by playing the speech signals through headphones. In this test, these ten 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 English TTS is shown in Table 1. and concatenated synthetic speech signals for three different sentences shown in Fig 5,Fig 6 and Fig7.

Test Sentences set	MOS
Sentence I	4.5
Sentence II	4.0
Sentence III	3.5

Table 2. Subjective Test Results

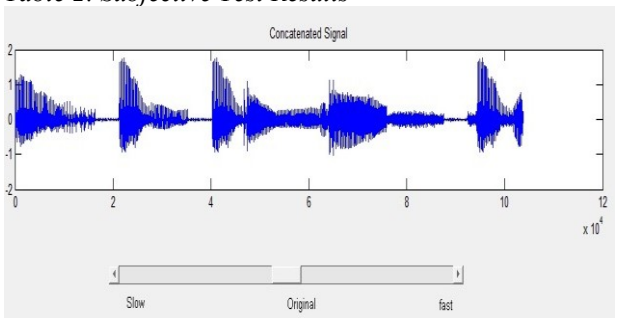

Fig.5. concatenated synthetic speech signal for "I am always happy"

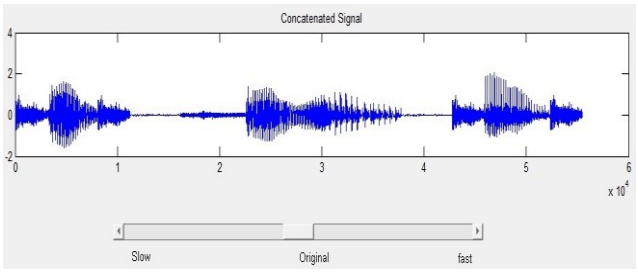

Fig.6. concatenated synthetic speech signal for phrase" Tit for Tat"

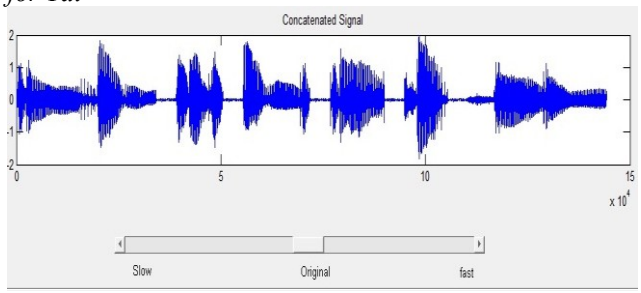

Fig.7.concatenated synthetic speech signal for quote "Dream big and dare to fail"

VI. COMPARISION WITH OTHER TTS SYSTEM

If we compare to other available TTS systems, there are some important differences between our system and available TTS systems. These are

1. This TTS is Simple flexible and user friendly TTS with least resources to provide various purposes.
2. Conventional unit section based TTS required 100 megabytes of memory; our TTS requires only 100 kilobytes of memory.
3. There is a fundamental difference between our system and any other talking machine in the sense that here we are interested in the real time operation.

VII. APPLICATIONS

It allows environmental barriers to be removed for people with a wide range of disabilities. The best use has been in the use of screen readers for citizens with visual impairment but TTS systems are now normally used by citizens with dyslexia and other understanding complicatedness as well as by preliterate kids [9]. They are also normally in use to aid those with severe speech injury usually through a dedicated voice output communication aid. Speech synthesis systems are also used in entertainment creations such as games and animations. In recent years, Text to Speech system used for disability and handicapped communication aids has become widely deployed in group Transit. Companies like Talking Signs and Text Speak Systems have pioneered solutions such as TTS for Digital Signage for the Blind that work via standard speakers and also radio receivers.

VIII. CONCLUSION

In this paper, we discussed the topics relevant to the development of TTS systems. We conducted MOS tests to evaluate the performance of speech synthesizer. This paper describes the successful completion of a simple text to speech translation by simple matrix operations. Thus this system is

ISSN: 2278 – 909X

very easy and efficient to implement unlike other methods which involve many complex algorithms and methods. The next step in improving this system would be implementing some machine learning algorithms in order to support generalization

REFERENCE

[1] S. D. Shirbahadurkar and D.S.Bormane "Subjective and Spectrogram Analysis of Speech Synthesizer for Marathi TTS Using Concatenative Synthesis." 2010 IEEE International Conference on Recent Trends in Information, Telecommunication and Comp ting

[2] Johnny Kanisha and G.Balakrishanan "Speech Transaction for Blinds Using Speech-Text-SpeechConversions" Advances in Computer Science and Information Technology Communications in Computer and Information Science Volume 131, 2011, pp 43-48

[3] Hamad, M." Arabic Text-To-Speech Synthesizer", Research and Development (SCOReD), 2011 IEEE Student Conference 9 978-1-4673-0099-5) on 19-20 Dec. 2011 409 - 414

[4] S.D.Shirbahadurkar and D.S.Bormane, (2009) "Marathi Language Speech Synthesizer Using Concatenative Synthesis Strategy (Spoken in Maharashtra, India)", Second International Conference on Machine Vision, pp. 181-185.

[5] Tapas Patra and Biplab Patra "Text to Speech Conversion with Phonematic Concatenation" International Journal of Electronics Communication and Computer Technology,Sept 2012,Issues 5, Volume 2

[6] Divya Bansal1," Punjabi Speech Synthesis System Using Htk" International Journal of Information Sciences and Techniques (IJIST) Vol.2, No.4, July 2012

[7] Aniruddha Sen "Pronucation Rules for Indian English Text to Speech System "Workshop on spoken language processing an ISCA supported event,Mumbai India , January 9-11 , 2003

[8]J.Sangeetha, S. Jothilakshmi and S.Sindhuja "Text to speech Synthesis for Tamil" International Journal of Emerging Technology and Advanced Engineering ISSN 2250-2459, Volume 3, Special Issue 1, January 2013

[9] Agnes Jacob and P.Mythili Research Article on Developing a Child Friendly Text-to-Speech System Hindawi Publishing Corporation Advances in Human-Computer Interaction, Volume 2008, Article ID 597971, 6 pages

ISSN: 2278 – 909X