International Journal of Agricultural Engineering, **Vol. 3 No. 1 (April, 2010) : 45-48**

Research Paper :

Speech synthesis

KOMAL SINGH **AND GEETA KAUSHIK**

Accepted : ***December, 2009***

ABSTRACT

Speech synthesis systems are often called text-to-speech (TTS) systems in reference to their ability to convert text into speech. However, systems exist that instead render speech synthesis to artificial production of human speech. A system used for this purpose is termed as a speech synthesizer and can be implemented in software or hardware. However, systems exist that instead render symbolic linguistic representations like phonetic transcriptions into speech. A text-to-speech system is composed of two parts: a front-end and a back-end. Broadly, the front-end takes input in the form of text and outputs a symbolic linguistic representation. The back-end takes the symbolic linguistic representation as input and outputs the synthesized speech waveform. TTS software can "read" text from a document, Web page or e-Book, generating synthesized speech through a computer's speakers. TTS can also convert text files into audio MP3 files that can then be transferred to a portable MP3 player or CD-ROM. This can save time by allowing the user to listen to reports or background materials while performing other tasks. TTS makes a critical difference to those with disabilities such as poor vision or visual dyslexia. People with speech loss can utilize specialized TTS programs to turn typed words into vocalization. TTS programs provide a valuable edge, particularly for learning new languages. This thesis aims to study the speech synthesis technology and to develop a cost effective, user friendly text to speech conversion system using Laboratory virtual instruments engineering workbench (LabVIEW) graphical programming language

See end of the article for authors' affiliations

Correspondence to:

GEETA KAUSHIK

Maharishi Markandeshwar University, Mullana, AMBALA (HARYANA) INDIA

Key words : Text-to-speech (TTS), Speech Synthesis

The aim of the speech synthesis is producing the human speech artificially either in software or hardware. The natural language text is converted into speech by text-to-speech (TTS) systems. These systems have been widely used as assistive technological tools for a long time. The pre-school kids, the people who have visual impairments or reading disabilities and the ones who suffer from severe speech impairment can get benefit from these systems. The news web sites that convert written news to audio content, entertainment productions such as games, cartoons, mobile tools, preparation of audio supplementary materials in various fields, automated question-answering systems, attaining certain information (price list, the weather forecasting report, etc.) and vocalizing e-mail, fax, sms, and daily journals for handicapped ones are only a limited number of items that can be listed as the typical application areas of TTS. The aim of the TTS is that the system converts all digital texts and printed texts into speech automatically. Commercial and non-commercial systems have been continuously developed and recent advances are promising for future applications.

Objectives:

Speech synthesis is the artificial production of human speech through the use of computer.

A very large set of different rules and their exceptions is needed to produce correct pronunciation and prosody for synthesized speech. The main objective of this report is to:

– Study the speech synthesis technology,

– Develop text to speech module using LabVIEW software

Synthesized speech can be produced by several different methods. All of these have some benefits and deficiencies. The methods are usually classified into three groups:

– Articulatory synthesis, which attempts to model the human speech production system directly.

– Formant synthesis, which models the pole frequencies of speech signal or transfer function of vocal tract based on source-filter-model.

– Concatenative synthesis, which uses different length prerecorded samples derived from natural speech.

The formant and concatenative methods are the most commonly used in present synthesis systems. The formant synthesis was dominant for long time, but today the concatenative method is becoming more and more popular.

METHODOLOGY

Concatenative synthesis is now the leading approach in speech synthesis, based on numbers of researchers pursuing that approach and numbers of commercial speech synthesizers using it. Many have also asserted that it delivers more natural speech than formant synthesis (though some would debate that assessment). Yet authors of The Bell Labs Approach state: "We use concatenative synthesis because that is currently the best available method to produce synthetic speech of consistently high quality.

Concatenative synthesis concerns the generation of speech from an input text. Concatenative synthesis can produce high-quality speech. Concatenative speech synthesis uses units of recorded speech, usually cut from full sentences. Commonly employed units are diphones (bracketing exactly one phone-to-phone transition, starting from the spectrally stable middle region of one phone to the spectrally stable middle region of the next phone), or demisyllables (comprising consonants and vowels).

A block diagram of a typical concatenative TTS system is shown in Fig. 1. The first block is the message text analysis module that takes ASCII message text and converts it to a series of phonetic symbols and prosody (fundamental frequency, duration and amplitude) targets. Input text is first analyzed and non-alphabetic symbols and abbreviations are expanded into full words. For example, in the sentence "Dr. Smith lives at 4305 Elm Dr.," the first "Dr." is transcribed as "Doctor," while the second one is transcribed as "Drive.

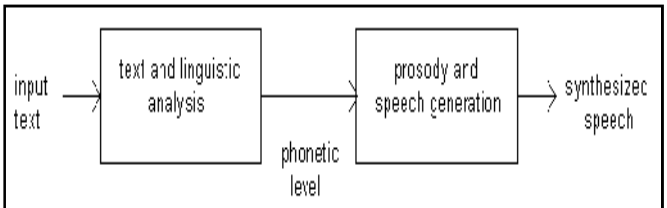

Fig. 1 : Simple text-to-speech synthesis procedure

The second block in Fig. 1 assembles the units according to the list of targets set by the front-end. It is this block that is responsible for the innovation towards much more natural sounding synthetic speech. Then the selected units are fed into a back-end speech synthesizer that generates the speech waveform for presentation to the listener.

In text to speech system the hardware requirements are very less. It requires only a good quality speaker for the production of sound signal. The software part is developed using the LabVIEW software.

Here, in text to speech module a text box is created so that user can write text which is to be converted into speech in .wav file format and creates a wave file named output .wav, which can be listen by using wave file player.

Flowchart for the text to wave file conversion is given in Fig. 2. This VI creates a wave format file named output .wav which consist the typed text converted into speech. The flow chart for capturing the speech signal is shown in the Fig. 2.

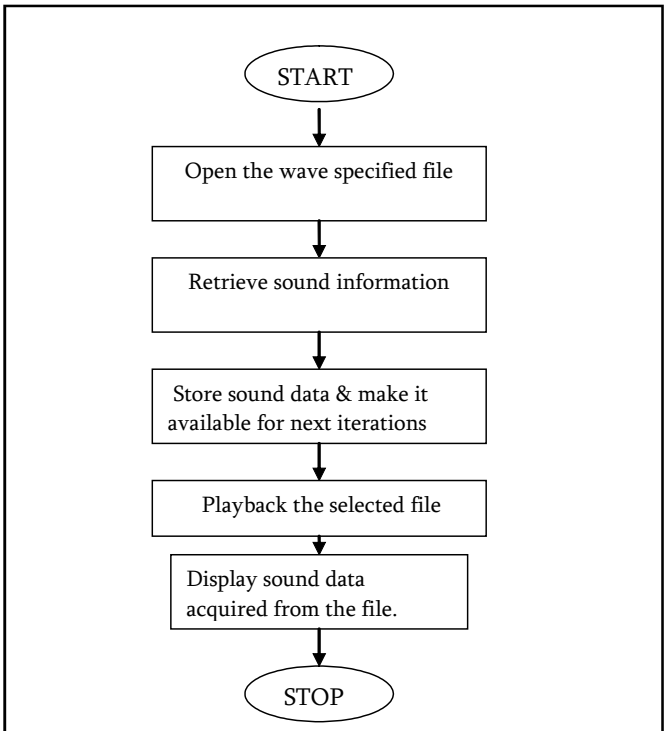

Fig. 2 : Flow chart for the text to wave file conversion

The various steps involved for recording the speech are as follows:

Step 1: Configure the input channel -

Sound format: Sound format specifies

– How the sound operation is set up (Mono or Stereo)

– Sets its playing rate (speed-11025, 22050 or 44100), and

– Sets up the sound as 8 or 16-bit sound.

Step 2: Initialize sound recording

Step 3: Start sound recording

Step 4: Store data in 20 ms frames until stop button is pressed.

Step 5: Returns resources used during recording to the system.

Step 6: Store sound data in wave file format.

RESULTS AND DISCUSSION

In this paper, a text to speech system using Concatenative method is developed using LabVIEW. This paper proposes a novel speech synthesis method to generate human-like natural speech. This method selects speech units from a large database, and concatenates them with or without modifying the prosody to generate synthetic speech. This method features highly human-like voice quality (Mizutani and Kagoshima, 2005). The

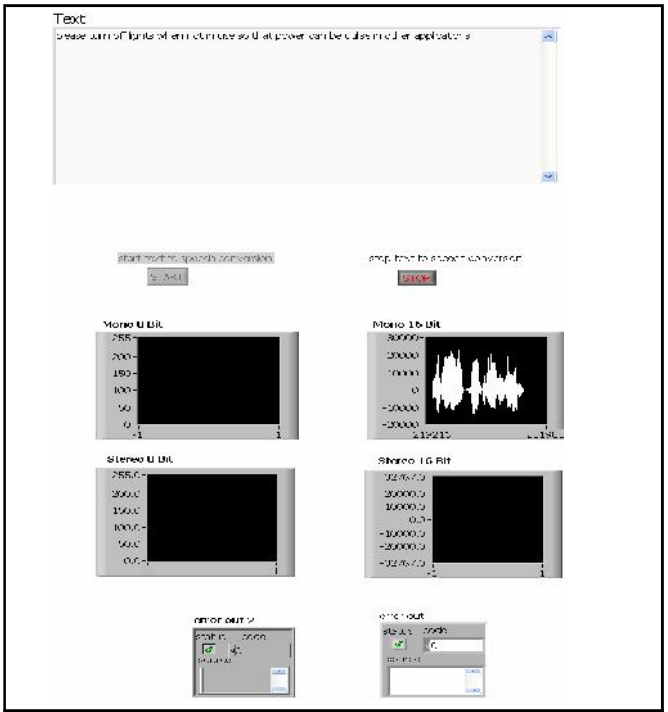

Fig. 3 : Front panel diagram of TEXT TO SPEECH .VI

availability of datalog files in LabVIEW makes it one of the most promising candidate for its usage as a database. Datalog files can access and manipulate data and complex data structures quickly and easily. It makes writing and reading much faster. TTS system is heavily dependent upon processing speed. Thus, faster is better. That's why we use LABVIEW than others.

The system developed in LabVIEW detects the user in almost real time with an accuracy of nearly 94% and is highly repeatable. As shown in Fig.3 we write some text in given text box and text to speech conversion is started. A wave file output. wave is created containing text converted into speech which can listen using wave file player as shown in Fig. 4.

The waveform will vary according to the different text typed in the text box and can be listened on the speaker. The wave form for "GOOD MORNING" is shown in Fig. 5.

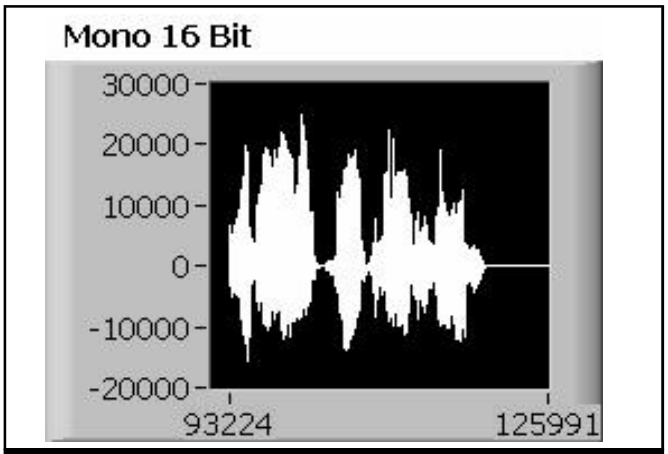

Fig. 4 : Use jof wave file player

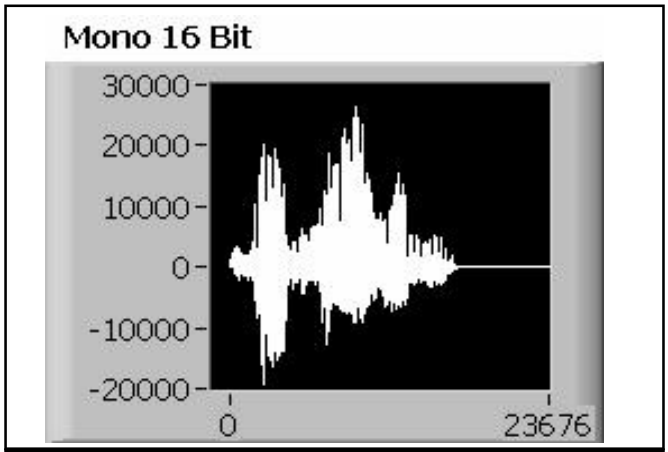

Fig. 5 : Waveform for "Good Morning"

Conclusion:

The LabVIEW tool allows us to implement the text to speech conversion .LabVIEW has its strong inbuilt Speech library to implement the text to speech conversion. Various methods and techniques are available for speech synthesis which have been discussed in this report. Here, in text to speech conversion VI, a text box is created so that user can write text which is to be converted into speech in .wav file format and creates a wave file named output .wav , which can be listen by using wave file player. The ACTIVE X sub pallet in Communication pallet is used to exchange data between applications. ActiveX technology provides a standard model for interapplication communication that different programming languages can implement on different platforms. Microsoft Speech Object Library (Fig. 3) has been used to build speech enabled applications, which retrieve the voice and audio output information available for computer. This library allows to select the voice and audio device one would like to use, enter the text to be read, and adjust the rate

and volume of the selected voice.

The application developed is user friendly, cost effective and gives the result in the real time. Moreover, the programme has the required flexibility to be modified easily if the need arises.

Authors' affiliations:
KOMAL SINGH, Department of Electronics and Communication Engineering, Maharishi Markandeshwar University, Mullana, AMBALA (HARYANA) INDIA

REFERENCES

Tatsuya, Mizutani and Takehiko, Kagoshima (2005). Corporate research & development center, Toshiba Corporation, Kawasaki-shi, 212–8582, Japan.

***