International Journal of Innovation and Scientific Research
ISSN 2351-8014 Vol. 17 No. 2 Aug. 2015, pp. 271-277
© 2015 Innovative Space of Scientific Research Journals
http://www.ijisr.issr-journals.org/

Speech to Text Conversion in Real-time

***Nuzhat Atiqua Nafis*[1] *and Md. Safaet Hossain*[2]**

[1]Department of Electronic and Telecommunication Engineering,
University of Development Alternative,
Dhaka, Bangladesh

[2]Department of Computer Science and Engineering,
University of Development Alternative,
Dhaka, Bangladesh

ABSTRACT: “Real time speech to text” can be defined as accurate conversion of words that represents uttered word instantly after speaking. Speech-to-text-conversion is a useful tool for integrating people with hearing impairments in oral communication settings, e. g. counseling interviews or conferences. However, the transfer of speech into written language in real time requires special techniques as it must be very fast and correct to be understandable. Our aim is to develop software that enhances the user's way of speech through correctness of pronunciation following the English phonetics. This software allows one to learn, judge and recognize their potential in English language. It also facilitates an extra add-on feature which nourishes the user's communication skills by an option of text to speech conversion also. The paper introduces and discusses different techniques for speech to text conversion and its process that described in complement with the options that are already in use. This paper presents a method to design a Text to Speech con version module by the use of Matlab. This method is simple to implement and involves much lesser use of memory spaces.

KEYWORDS: speech to text, properties of speech, Speech to text system, another way of data entry, makes communication easier for handicapped users, natural language processing.

1 INTRODUCTION

For past several decades, designers have used or processed speech for different applications. Speech recognition reduces the difficulties and problems caused by other communication methods. In the past speech has not been used much in the field of electronics and computers. However, with modern processes, algorithms, and methods we can process speech signals easily and use it in our desirable fields. Our speech-to-text engine directly converts speech to text. It can complement the idea giving users a different choice for data entry. Our speech-to-text engine can also provide data entry options for blind, deaf, or physically handicapped users. Text-to-speech convention transforms linguistic information stored as data or text into speech. It is widely used in audio reading devices for blind people now a days [1].In the last few years however; the use of text-to-speech conversion technology has grown far beyond the disabled community to become a major adjunct to the rapidly growing use of digital voice storage for voice mail and voice response systems. Also developments in Speech synthesis technology for various languages have already taken place [2] [3]. Many speech synthesizers using complex neural networks have also been designed [4]. In the bigger picture, the module can open up a window of opportunities for the less privileged paving the way for a plethora of employment opportunities for them in the job sector. It can also play a defining role in establishing communication of the blind if it is incorporated into mobile phones (so that text messages could be converted into speech). [5] [6]. The existing system deals with various dictionaries, which implements dictation of words with correct pronunciation. It supports the operation, only for a set of words, which are available in the dictionary. The availability of such software doesn’t eradicate the problem, which is being discussed in the picture. The system speaks out the selected word, which the user wishes to listen to. The current system is focused more on polishing the pronunciation from better to best and not focused to bringing up someone from nothing to best.

2 Properties Of Speech

The most natural way to communicate for human beings is through words. Human respiratory and articulator systems which consist of different organs and muscles generate speech [8][9]. Coordinated action of these speech production organs creates speech. The speech or voice can be categorized in many ways. In general, the following ways are mainly analyzed: acoustic, phonetic, phonological, morphological, syntactic, semantic and pragmatic. These levels are used or being processed in the way of converting text to speech. Furthermore, communication in spoken language can be divided as linguistic and paralinguistic. Paralinguistic information is considered as information about the speaker, way how words are spoken, and the factors during speech like breathing, speed, intonation, hesitation etc. And linguistic is considered as the actual meaning of word. Mainly Paralinguistic is the way of communication and linguistic is the information conveyed during communication.

3 Theoretical Approach

Over the past 20 years several developers and designers has improved the way of converting speech to text in real time [15]. It is their hard work that we are able to convert speech to text. The developments are done by improving technologies, computer systems and communication ways. These parallel developments led the way to the applications we use today for converting speech into text.

Currently, two major options are available for providing real-time speech-to-text services:

1. Computer assisted note taking (CAN).
2. Communication access (or computer aided) real-time translation

3.1 Difference Between Methods

There is a lot of difference between these two methods:

1. In their process of generating speech to text [11] in real time [12].
2. With respect to the circumstances under which the methods can be properly used and
3. With respect to the amount of training which is will help to convert speech to text successfully.

3.2 Speech Preprocessing

We used MATLAB software’s sound recorder tools for speech processing and also use those steps:-

i. The system must identify useful or significant samples from the speech signal. To accomplish this goal, the system divides the speech samples into overlapped frames

ii. The system checks the frames for voice activity using endpoint detection and energy threshold calculations.

iii. The speech samples are passed through a pre-emphasis filter.

iv. The system performs autocorrelation analysis on each frame.

We implemented these steps with C# programming, which was executed based on the algorithms. Although we used the same C# programming for training and recognition, the C# code executes on a PC during training. Then we downloaded the Nios II software. This is embedded software. It will supply the Hidden Markov Model (HMM). It will also supply the dedicated tools for Quartus II software. With the help of the journal paper we found the names of these programs and their work.

We could have performed software design and simulation using the Quartus II software and use SOPC [6] Builder to create the system from readily available, easy-to-use components. With the Nios II IDE, we can easily create application software for the Nios II processor with the intuitive click-to-run IDE interface. SOPC Builder’s built-in support for interfaces and the easy programming interface provided by the Nios II software application layer make the Nios II processor for implementing our on-line speech-to-text system.

We started doing the work in order to build our project with the journal, but the Quartus II software needs to be programmed in .qsf language. It was a problem. So we started digging our way up.

4 Speech To Text System

Our speech-to text system converts speech to text from instantly given voice. This system does not synthesize the quality of recorded human speech. There are different technologies suitable for different applications.

Basically, there is no simple metric that could be applied to any STT [Speech to Text] system and which would give a clear concept of the overall quality of any system. The main reason is that the STT system should not be assessed in isolated place, but it should be evaluated for their respective uses. There are many uses for STT systems, so they should be given to their exact destinations.

4.1 Building Process

In the process of completing our project, we have to go through certain processes. Those processes are described below:

i. Recording audio & converting it into .wav format

ii. Processing that .wav file

iii. Storing it in a file

iv. Making software to compare the audio with other audio files with inserted voice and recognize it

v. Making a program to show the voice files in text format.

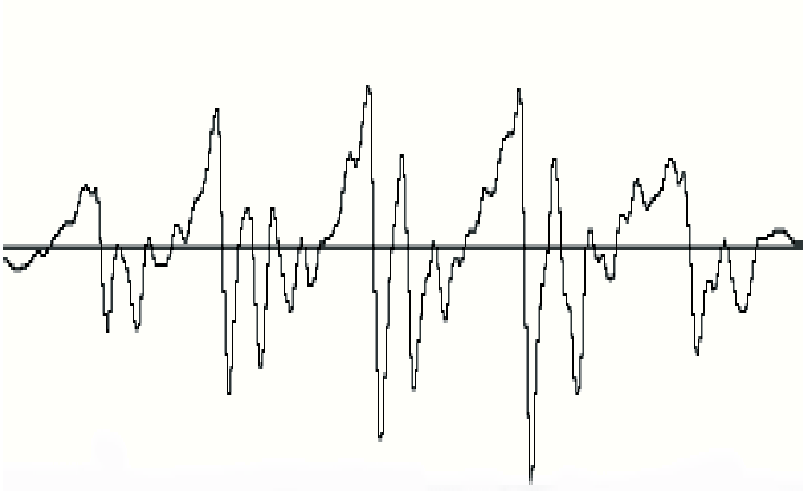

Fig. 1. Part of the waveform sound done with Matlab software

4.1.1 Requirements

- Personal Computer (Desktop/laptop)
- Matlab Version 11
- Visual Studio 2010
- Audio Recorder Toolbox
- Microphone & Headset

4.2 Work Process

Our speech-to-text system directly acquires and converts speech to text. We built the project using Microsoft Visual Studio. But to use the direct conversion we have to add Speech SDK 5.0. We downloaded this software from MSDN Library. Then we installed that program. We added this software with Visual studio. Then we started building the program. We took a simple windows form application.

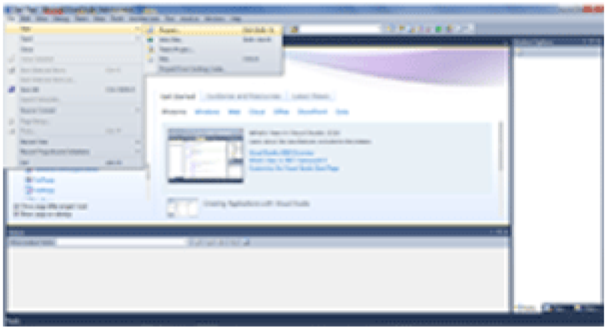

Fig. 2. Starting a new project with visual studio 10

Then we had to add reference. We added .speechlib as a reference. We also added system. sound for building the project correctly.

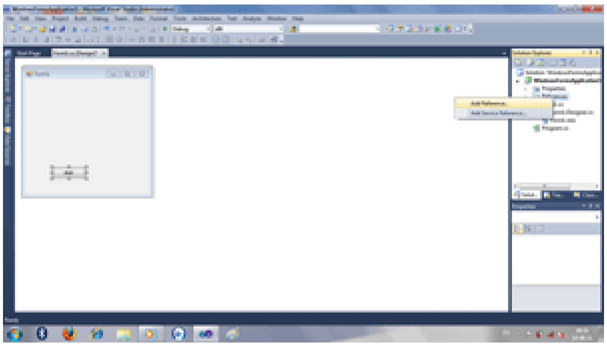

Fig. 3. adding a reference

After that we have added a button to the form. The button is called 'Start dictation'.

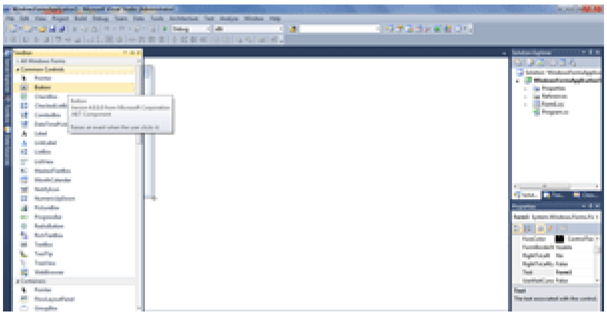

Fig. 4. adding a button

We did the code for the project in C#. The entire project is based on C# on Microsoft Visual Studio 10. We have added the Dictation format from Speech SDK 5.1. The project is based on a voice database of Microsoft.

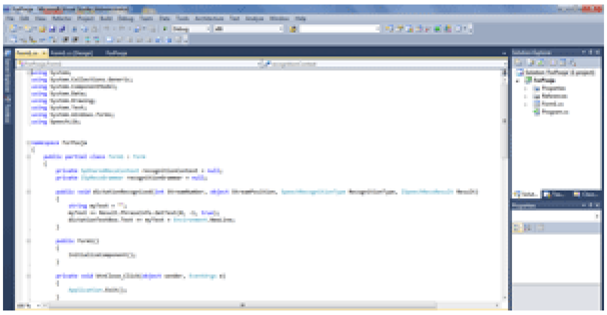

Fig. 5. coding of the project

4.3 Implementation

The project implements a speech-to-text system using isolated word recognition with a vocabulary of limited words (recognized with SDK 5.1) and statistical modeling (HMM) for machine speech recognition. In the training phase, the uttered digits are recorded using 16-bit pulse code modulation (PCM) with a sampling rate of 8 KHz and saved as a wave file using sound recorder software. We use the MATLAB software's wavread command to convert the .wav files to speech samples.

Generally, a speech signal consists of noise-speech-noise. The detection of actual speech in the given samples is important. We divided the speech signal into frames of 450 samples each with an overlap of 300 samples, i.e., two-thirds of a frame length. The speech is separated from the pauses using voice activity detection (VAD) techniques.

The system performs speech analysis using the linear predictive coding (LPC) method. From the LPC coefficients we get the weighted cepstral coefficients and cepstral time derivatives, which form the feature vector for a frame. Then, the system performs vector quantization using a vector codebook. The resulting vectors form the observation sequence. For each word in the vocabulary, the system builds an HMM model and trains the model during the training phase is performed using PC-based C programs.

5 Result

We took samples from various users by making them pronounce the same word. The following table shows the result:

Table 1. Table of Words

Words	**User 1**	**User 2**	**User 3**	**User 4**
Hello	Hello	Haul	Hello	Hello
Excuse me	Seem	Yes me	Excuse me	Excuse me
Thank you	Thank you	Thank you	Thank to	Thank you
One	One	One	One	One
Accuracy	**75%**	**50%**	**75%**	**100%**

5.1 Performance Measurement

The performance of the project is very good. It recognizes the words with its added database. If the words match then it shows the output. Its accuracy is 75%

5.1.1 Advantage

We do have a few advantages comparing to the previous method described in the journal. They are:

i. Timing: Our timing is better. [Because we are converting speech to text instantly, within 2-3 sec]

ii. Costing: Our costing is much less. [Because we are doing it in our personal computers and using only 2 softwares]

iii. Hassle of using lot things: We have only used two softwares- Matlab & Visual Studio 10

5.1.2 Disadvantage

i. The disadvantage is that our project will only run in those computers's which has visual studio in it.

ii. Another disadvantage is with our accent. The SDK only recognizes American accent. So it sometimes miswrites our voice messages.

5.2 Target Users

The design can be used for various applications. The basic system can be used in applications such as:

- Interactive voice response system (IVRS) [7]
- Voice-dialing in mobile phones and telephones [10]
- Hands-free dialing in wireless Bluetooth headsets

- ■ PIN and numeric password entry modules
- ■ Automated teller machines (ATMs)
- ■ Data entry work
- ■ In classroom works for disabled students

6 DISCUSSION

A low cost, fully functional speech to text converter meets the need of converting voice into text. It is very good and useful. As Bangladesh [16] is a developing country, here needs to develop in software side. For developing the sector you need to be eager and research minded. If you can promote this, it will be very helpful and inspired. Neither this project is costly nor critical, so it is available.

7 FUTURE WORK

Our project is a fully functional complete project. Our project can be used as various School/College/University labs. It can be also used for various research projects. We can improve the performance of the project by training our computers with our voice. By this method the computer will come to know our voice, the way we speak, our accent. This will help to make this project's performance better. Another scope of changing is that this project can be built for mobile [9]. So it will be easier for people to use this project.

REFERENCES

[1] Leija, L.Santiago, S.Alvarado, C., "A System of text reading and translation to voice for blind persons," Engineering in Medicine and Biology Society, 1996.

[2] R. Sproat, J. Hu, H. Chen, "Emu: An e-mail preprocessor for text-to-speech,‟‟ Proc. IEEE Workshop on Multimedia Signal Proc., pp. 239–244, Dec. 1998.

[3] C.H. Wu and J.H. Chen, "Speech activated telephony e-mail reader (SATER) based on speaker verification and text-to-speech conversion,‟‟ IEEE Trans. Consumer Electronics, vol. 43, no. 3, pp. 707-716, Aug. 1997.

[4] Ainsworth, W.,"A System for converting English text into speech, "Audio and Electroacoustics, IEEE Transactions, vol.21, no.3, pp. 288-290, Jun 1973.

[5] Nipon Chinathimatmongkhon, Atiwong Suchato,Proadpran Punyabukkana, "Implementing Thai text-to-speech synthesis for hand held devices", Proc Of ECTI-CON 2008.

[6] M.T. Bala Murugan and M. Balaji , "SOPC-Based Speech-to-Text Conversion" National Institute of Technology, Trichy.

[7] Ki-Yeol Seo, Se-Woong Oh, Sang-Hyun Suh, and Gyei-Kark Park, "Intelligent Steering Control System Based on Voice Instructions" vol. 5, no. 5, pp. 539-546, October 2007.

[8] Sangam P. Borkar, "Text to speech system for Konkani (GOAN) language" Rajarambapu Institute of Technology Sakharale, Islampur, Maharashtra, India.

[9] Kimmu Parssinen, "Multilingual Text to Speech system for Mobile Device" University of technology April 2007

[10] Julie Ngan, Joseph Picone, "Issues in generating pronouncing pronunciation dictionaries for voice interfaces to spatial databases" Institute for Signal and Information Processing, Mississippi State University.

[11] Susanne Wagner (Halle), "Intralingua speech-to-text-conversion in real-time: Challenges and Opportunities" Muttra 2005 – Challenges of Multidimensional Translation

[12] Yee-Ling Lu; Mak, Man-Wai; Wan-Chi Siu,, "Application of a fast real time recurrent learning algorithm to text-to-phoneme conversion," Neural Networks, 1995. Proceedings., IEEE International Conference on , vol.5, no., pp.2853,2857 vol.5, Nov/Dec 1995.

[13] Decadt, Bart; Duchateau, Jacques; Daelemans, Walter; Wambacq, P., "Phoneme-to-grapheme conversion for out-of-vocabulary words in large vocabulary speech recognition," Automatic Speech Recognition and Understanding, 2001. ASRU '01. IEEE Workshop on , vol., no., pp.413,416, 2001.

[14] Penagarikano, M.; Bordel, G., "Speech-to-text translation by a non-word lexical unit based system, "Signal Processing and Its Applications, 1999. ISSPA '99. Proceedings of the Fifth International Symposium on , vol.1, no., pp.111,114 vol.1, 1999.

[15] Olabe, J. C.; Santos, A.; Martinez, R.; Munoz, E.; Martinez, M.; Quilis, A.; Bernstein, J., "Real time text-to-speech conversion system for spanish," Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '84. , vol.9, no., pp.85,87, Mar 1984.

[16] Sultana, S.; Akhand, M. A H; Das, P.K.; Hafizur Rahman, M.M., "Bangla Speech-to-Text conversion using SAPI," Computer and Communication Engineering (ICCCE), 2012 International Conference on , vol., no., pp.385,390, 3-5 July 2012.

[17] F.; Moulines, E., "Text-to-speech algorithms based on FFT synthesis," Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on , vol., no., pp.667,670 vol.1, 11-14 Apr

[18] Stuckless, Ross (1999): “Recognition Means More Than Just Getting the Words Right”. Speech Technology Oct/Nov 1999, 30.

[19] Wagner, Susanne & Kämpf de Salazar, Christiane (2004): “Einfache Texte – Grundlage für barrierefreie Kommunikation“. In Schlenker-Schulte, Christa (ed.): Barrierefreie Information und Kommuniaktion: Hören - Sehen – Verstehen in Arbeit und Alltag. WBL. Villingen-Schwenningen: Neckar-Verlag.

[20] Wagner, Susanne & Prinz, Ronald & Bierstedt, Christoph & Brodowsky, Walter & Schlenker-Schulte, Christa (2004): “Accessible Multimedia: status-quo, trends and visions”. IT - Information Technology 6. 346-352.