Hardware Approach of Text-to-Speech in Embedded Applications: Work in Progress

Gordana Laštovička-Medin, Itana Bubanja
Faculty of Science and Mathematics
University of Montenegro
Podgorica, Montenegro

Abstract—**This paper presents work in progress. Here, we describe our ongoing experience teaching embedded systems to physics students who've been given access to the Arduino platform and its open source community. Our chosen topic, Design and Applications of Embedded Systems for Speech Processing, was researched as part of the Basic Measurement in Physics course of the Faculty of Science and Mathematics at the University of Montenegro. During the student research the SpeakJet sound synthesizer was explored. Building embedded systems has enormous potential for developing students' skills and cultivating a culture of thinking and participating**

Keywords-speech processing, speech synthesis, SpeakJet, processor TTS, Arduino

I. INTRODUCTION

Speech processing, especially audio manipulation and sound processing is commonly performed by many everyday electronic devices. Examples of such devices are are digital voice recorders, speaking GPS receivers, and many others. In general, the speech processing capabilities that can be added to an electronic device are voice recording, voice playback, text-to-speech (TTS) synthesis and speech recognition (SR). Voice recording and voice playback are used in digital voice recorders to store speech in non-volatile memory and then replay it at a later time. TTS involves reading a written text and converting it into spoken words that can be played through speakers. People with reading or visual difficulties may find such systems extremely useful.

TTS synthesis transforms any linguistic information stored as data or text into speech. We can make robots speak by simply recording human speech, and playing it back when needed. True speech synthesis, however, is to allow robots to generate boundless speech output- in other words to let them speak their mind. TTS is true speech synthesis. It is the synthesis of speech based on unrestricted text input. There are three main classes of TTS synthesis: articulatory, formant, and concatenative [1]. Articulatory synthesis is based on a complete 3D model of the human speech apparatus. It uses acoustic parameters extracted from the model to synthesize speech. Articulatory synthesis is the most powerful process, but also the most complex. It requires analyzing Magnetic Resonance Imaging (MRI) scans of speech production. It scores high in intelligibility, but low in naturalness. Formant synthesis uses a black-box modeling approach to speech production It analyzes the end transfer function of the vocal tract, rather than the way it is made. Formants are the vocal tract's resonant frequencies. They give the phonetic character to speech sounds. The voice of Stephen Hawking is an example of formant synthesis. It shows high intelligibility but low naturalness. Concatenative synthesis does not seek to model speech production. It uses a database of prerecorded segments of natural speech that it concatenates one after the other. This approach gives the synthetic speech a very natural sound.

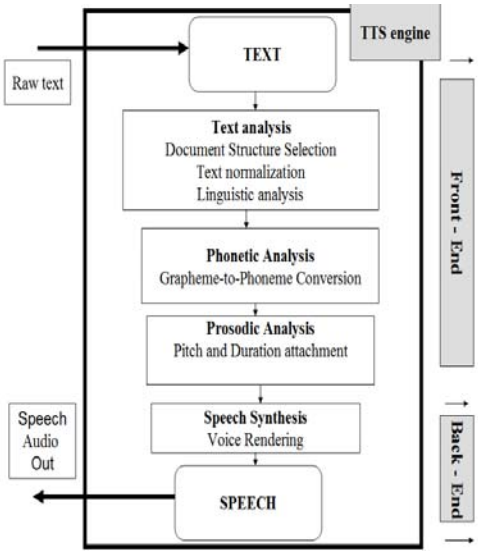

Figure 1. Block diagram of a general text-to-speech system. The figure has been adopted from [1].

A text-to-speech system (or "engine") is composed of two parts: front end and back end. The front end has two major tasks, as illustrated in Fig. 1 [1]. Firstly, it converts raw text containing symbols, such as numbers and mabbreviations, into the equivalent of written words. This process is often called text normalization, pre-processing, or tokenization. The front end then assigns phonetic transcriptions to each word, and divides and marks the text into prosodic units, like phrases, clauses, and sentences. The process of assigning phonetic transcriptions to words is called text-to-phoneme or grapheme-to-phoneme conversion. Phonetic transcriptions and prosodic information together make up the symbolic linguistic representation that is output by the front end. The back end, often referred to as the synthesizer, then converts the symbolic linguistic representation into sound. For more details we refer to [2,3,4].

TTS in embedded systems can be analyzed from several points of view: as TTS integrated circuits, TTS modules, TTS across embedded operating systems and TTS software applications for embedded devices. Details can be found in [5]. In this article we are interested in TTS integrated circuits (ICs). We use a SpeakJet sound synthesizer chip and the TTS256 Text-to-Speech chip for SpeakJet, in order to create a text-to-speech solution. The work is under progress.

II. SPEKAJET: ALLOPHONE BASED SPEECH SYNTHESIS

The central hardware component of this project is the Magnevation's Speakjet which is a 20-pin IC [6]. It is designed to add speech and audio to embedded microcontroller applications. The chip is self-contained and requires just an external +5V supply and a speaker for its operation. A mathematical sound algorithm is used to control its five channel internal sound synthesizer to generate vocabulary speech synthesis and complex sound generation. The chip is low cost and is aimed primarily at the hobby market. The Speakjet is programmed with 72 speech elements, 43 sound effects and 12 DTMF touch tones. In addition, sound effects such as the pitch, rate, bend and volume can be controlled. The chip can easily be controlled from a microcontroller.

The SpeakJet uses a technique called allophone based speech synthesis in order to create the sound that we interpret as intelligible speech. A string of letters spelled out as "hello world" can not be sent to the SpeakJet because the way the words are written down in English and the way they are spoken is very different. Written English text consists of a series of letters but spoken text consists of a series of phonemes The smallest meaningful unit of sound in human speech is called a "phoneme". Phonemes, in turn are represented by allophones which are sets of multiple possible spoken sounds used to pronounce a single phoneme. To be able to generate intelligible speech from an allophone based synthesizer it is important to understand the difference between letters and allophones. There are 26 letters in the English alphabet but hundreds of allophones. English language is not spoken phonetically since subconsciously speakers apply various conventions that change the sound represented by particular letters based the context surrounding the word or sentence. For example, the letter "e" may be short as in "set" or it can be long, as in the first e in "concrete". Contrary, the most southern Slavic languages (Montenegrin, Bosnian, Serbian) are spelt phonetically. The writing Serbian system does not take into account allophones while as we saw before the allophones play a critical key in naturalness of English synthesized speech.

Figure 2 shows block diagrams of the system designed to control, manipulate and program the SpeakJet. The circuit diagram is displayed in Fig. 3. The aim was to program chip to be used in talking puppet, thus there are 3 tackle switches (S1,S2,S3) in order to provide chip activation by pressing toy\s hands on three different places. Besides the SpeakJet, two additional integrated circuits were used to build the complete electronic circuit: an LM386 low-voltage audio power amplifier and a MAX232 driver/receiver.

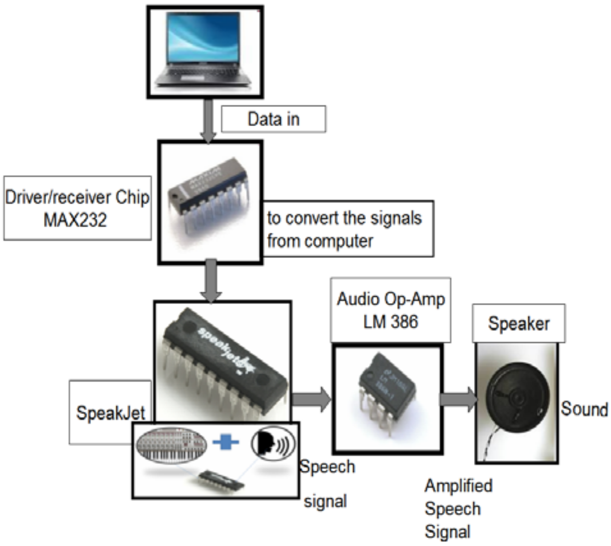

Figure 2. Block diagram designed used to program SpeakJet.

The MAX232N-1 integrated circuit was used to convert the signal from a serial port to signals suitable for use in digital logic circuits. The created signal was transferred using 6 phono jacks and phono plugs. By pressing one of three tactile switches connected to the phono plugs, an electric circuit was created and the SpeakJet was "ready" for uploading. A tactile switch also known as a momentary button or push-to-make switch, is commonly used for inputs and controller resets. These types of switches create a temporary electrical connection when pressed. One pin is supplied with +5 volts and the other pin is grounded. The LM386N-1 audio amplifier takes the electrical signal generated by the SpeakJet when we push and release one of the tactile switches (S1 S2, S3) and then amplifies the electrical signal to create enough power to drive the speaker. Switches S1, S2, and S3 were used to control the voltage on SpeakJet's pins 2, 4, and 7. The switches are normally open, which means each pin is connected to ground. When we push one of the switches, the voltage on the corresponding pin raises to +4.5 volts. This means that electric circuit is created and SpeakJet is "ready" for software uploading. for use in digital logic circuits.

The hardware system accompanying the circuit diagram is displayed in Fig. 4. A breadboard was used for circuit prototyping. Fig. 9 shows the same design as presented in Fig. 4 but with added phono plugs and phono jacks. A useful tool from Magnevatron is a Windows program called Phrase-A-Lator which has a great dictionary of word-to-allophone translations, and also has the ability to pump information directly at a Speakjet connected to a PC via a serial port. This is very useful because in this way phrases are directly programmed into the EEPROM, and can be tested for sound. Using the Phrase-A-Lator, we can convert the text/allophone string into the proper codes to be sent to the Speakjet chip, and apply those to our Arduino code (see next chapter).

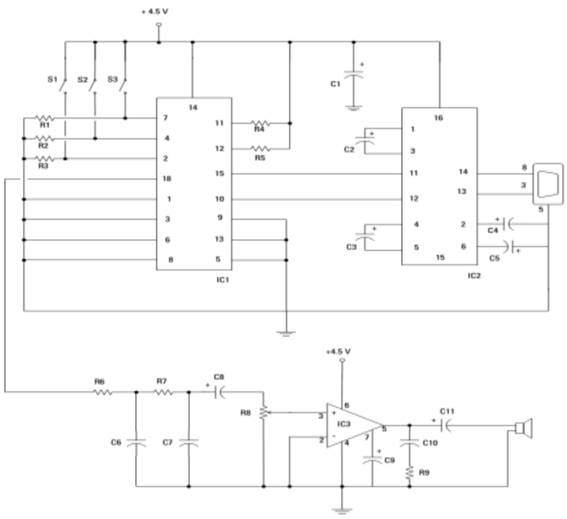

Figure 3. A circuit diagram with SpeakJet, Max232 and audio op-amplifier LM386 to program SpeakJet.

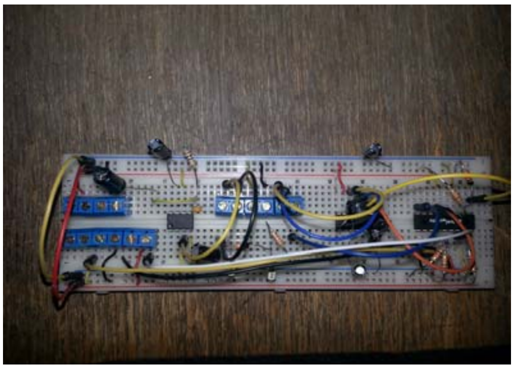

Figure 4. Breadboard SpeakJet prototyping

Figure 5. Breadboard as shown in Figure 4 but with added phono plugs and phono jacks.

III. TEXT-SPEECH SOLUTION WITH SPEAKJET AND TTS256

Speech processing with the IC SpeakJet works well for a limited vocabulary, but if an application requires unlimited or arbitrary speech output, the TTS chip does the work of translating any English text into the allophenes that the SpeakJet understands. The TTS256 chip [6,7,8], which contains a dictionary of words-to-allophones converts English text into a sequence of phonemes. This chip is a companion to the SpeakJet and comes with a built-in 600-rule database to convert English text to phoneme codes. Speech can easily be generated from ASCII text in microcontroller-based embedded applications, making the chip extremely easy to use in applications where speech generation is required. The TTS256 is controlled from its serial port and, thus, it is compatible with any microcontroller with such a port. Then a SpeakJet chip converts the phonemes into sound. The problem is that the Magnevation software can not communicate directly with this chip, since the communication is with Arduino itself. All that is needed is a simple little host program on the Arduino to redirect information to the Speakjet.

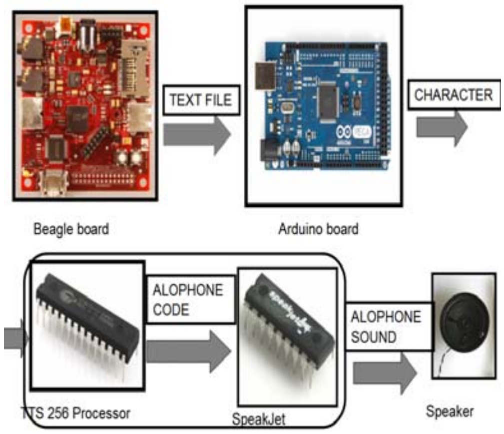

Figure 6. Block diagram of speech synthesizer. It consists of BeagleBoard, Arduino, TTS256, SpeakJet and Speaker

Figure 6 shows a block diagram of a speech synthesizer. It consists of the BeagleBoard, Arduino, the TTS256, the SpeakJet and a Speaker. The Arduino was used to control TTS256 while the Beagle Board [9] was used as a single-board computer which is is low-power, open-source hardware [9]. The TTS-Speakjet breadboard prototyping is displayed in Figure 7.

In what follows we will show how the Magnevation software works in order to program chip. The Magnevation software opens and closes the serial port every time we send information down. Arduino's default setup will run the following process: when the USB-Serial connection is opened up by the host, it institutes a chip reset, and the program restarts. Figure 8 shows a couple of Phrase-A-Lator screenshots. When the phrase is completed (with the help of "Say it"), we have to select "View Codes" to get the numerical allophone sequence and then to paste it into the Arduino code. Alternatively to the setup shown in Fig. 9 the VoiceBox shield can be used. It contains a SpeakJet chip while the TTS chip can be soldered directly onto the VoiceBox shield, as shown in Fig. 9.

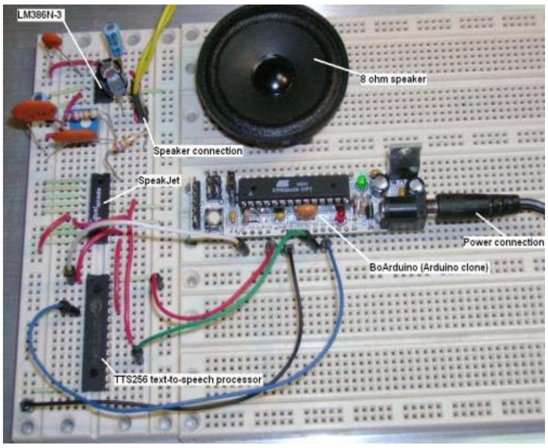

Figure 7. Breadboarding design of electronic circuit containing of TTS256, Speakjet and Speaker

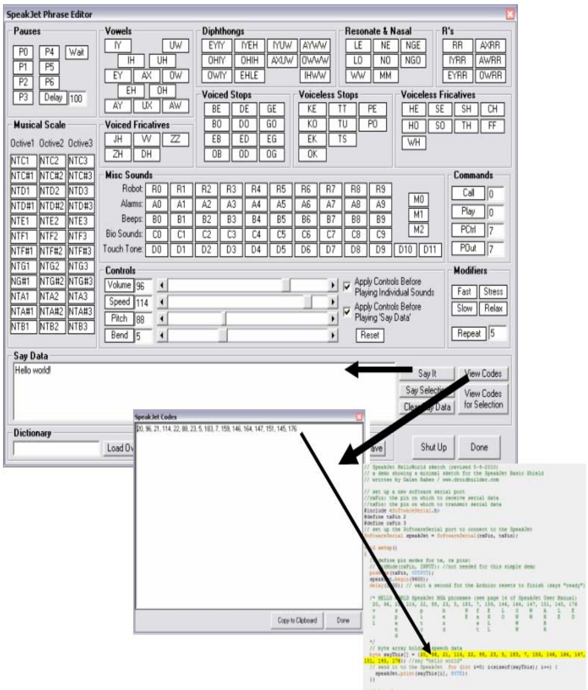

Figure 8. PhraseALator's control panels.

IV. FUTURE WORK AND DISCUSSIONS

Our research towards speech processing is at its early stage. An attempt had been made in order to implement and test the text-to-speech solution. The hardware design is based on a TTS256 chip, a sound synthesizer SpeakJet and an open source platform. Paper also presents an pedagogical aspect of TTS engine where students explored the hardware approach to TTS and how to program SpeakJet. Currently we extended our research towards the design of the audio hardware interfaces in order to visualize and manipulate the audio signal. The issues we are faced are the choice of the signal encoding schemes, methods for signal visualization and the selection of deployment platform. Additionally, the TTS-interface can also be used creatively to create dynamic facial animation. The project has potential to be additionally extended towards designing the multimodal platforms in order to study the various combinations of audio-visual speech processing, including real-time lip motion analysis, real-time synthesis of models of the lips and of the face, audiovisual speech recognition of isolated words, and text-to-audio-visual speech synthesis in Montenegrin

.

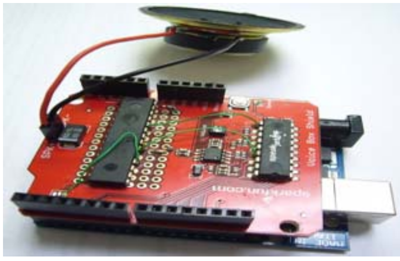

Figure 9. VoiceBox with SpeakJet and TTS256

REFERENCES

[1] X. Huang, A. Acero, H.-W. Hon, Spoken Language Processing, Prentice Hall PTR, 2001

[2] Tanja Schultz (Edit.), Katrin Kirchhoff (Edit.), Multilingual Speech Processing, Elsevier, 2006

[3] Thierry Dutoit , An Introduction to Text-to-Speech Synthesis, Published by Kluwer Academic Publishers, ISBN 0-7923-4498-7, The Netherlands, 1997

[4] Vincent J. van Heuven, Louis C. W. Pols, Analysis and Synhesis of speech: Strategic Research towards High-quality text-speech-generation, Published by Mounton de Grayter, 1993

[5] Shrikanth Narayanan, Abeer Alwan, Text to Speech Synthesis: New Paradigms and Advances, Published by Prentice Hall. Part of the IMSC Press Multimedia Series, 2004-ISBN-13: 978-0-13-145661-7

[6] http://magnevation.com/software.htm

[7] htto::/www.magnevation.com/pdfs/speakjetusermanual.pdf

[8] www.speechchips.com

[9] http://beagleboard.org/