ISSN 2277-3061

Multilingual Text to Speech in embedded systems using RC8660

Azadeh Nazemi, Iain Murray & David A. McMeekin

Department of Electrical and Computer Engineering, Curtin University, Perth, WA, Australia

Azadeh.nazemi@postgrad.curtin.edu.au

Department of Electrical and Computer Engineering, Curtin University, Perth, WA, Australia

I.murray@curtin.edu.au

Department of Spatial Sciences,Curtin University, Perth, WA, Australia

D.McMeekin@curtin.edu.au

ABSTRACT

Most multilingual Test to Speech (TTS) systems are software applications which allow people with visual impairments or reading disabilities to listen the written material using computer. This paper describes an approach to make a multilingual TTS and embed it into the portable, low cost, and standalone embedded system to access and read electronic documents particularly in developing countries. There are several TTS such as Doubletalk, DECtalk, and Dolphin available in market, also there are some products using TTS such as Talking OCR, Bill Reader and Intel Reader, which are not affordable or multilingual. To design this system OMAP3530 an application processor board is considered as the hardware platform to process the language-independent parts of the application and RC8660 used as an integrated TTS processor**.**

Indexing terms/Keywords

Lexical to Markup Framework (LMF), Prosody, Romanization, Speech synthesizer, Text to Speech (TTS).

Council for Innovative Research

Peer Review Research Publishing System

Journal: INTERNATIONAL JOURNAL OF COMPUTERS AND TECHNOLOGY

Vol. 13, No. 4

editorijctonline@gmail.com

www.cirworld.org/journals

INTRODUCTION

Text to Speech systemsuse in packages for ATMs, Kiosks, Vending, Ticketing and Banking systems that require audio for the sight impaired to meet local requirements.

Text to Speech at bus stops & rail platforms providesreal-time information for passengers specifically vision impaired

TTS systems vary in their reliability and intelligence and can be implemented in software or hardware. Some TTS give the user real time control of the speech signal, including pitch, volume, tone, speed, expression, and articulation. The main target of a text to speech system is producing natural sounding speech from the input plain text of ASCII or UNICODE. The TTS system generally has two major modules:

1)Text analysis module

2)Synthetic speech producing module [1]

The TTS systems must first convert the input text into its corresponding linguistic or phonetic representations and then produce the sounds corresponding to those representations.

TTS STRUCTURE

A. Module 1

The conversion in the first module is highly language dependent. In this stage, the sentences in the text are divided into words, numbers, abbreviations and acronyms.

Phonetic realizations of the segments are dependent on context both within words and across word boundaries. Determination of phonetic transcription of the text is performed using dictionary-based or rule-based methods. In this module, access to the complete databases and recourses is a decisive factor to achieve high quality synthesized speech [2].

B. Module 2

Synthesizing of speech could be composed by concatenating pre-recorded samples derived from natural speech. Due to the huge number of words and phrase, recording and storing all words and concatenating the words in the given text to produce, the natural corresponding speech is not feasible [3].

Text is synthesized by selecting appropriate units from a speech database and concatenating them. The most effective factors in the quality of synthesized speech are fundamental frequency, speed of speech and the availability of appropriate units with proper prosodic features in the database [4]. Figure 1 illustrates the block diagram of the TTS system.

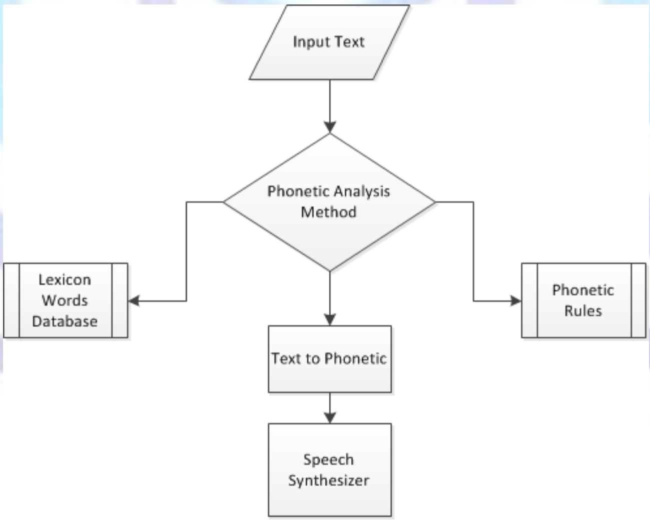

Fig1. TTS functional block diagram

PROSODY AND PROSODIC MARKUP

Prosody is rhythm, stress and intonation of speech, which conveys aspects of meaning and structure. It is not implicit in the segmental content of utterances. It operates on longer linguistic units more than the basic speech units do. It is based on:

- Pauses between the two words
- Pitch
- Phoneme duration and time
- Relative amplitude or volume

- Stress[5][6]

In most cases, input text does not contain explicit information about the desired prosody so prosodic realization is a challenging task. Prosodic phrasing involves finding meaningful prosodic phrases, increases the understandability of synthesized speech. It can be possible by creating prosodic boundaries at explicit identifiers like punctuation marks and grammatical words. The previous researches undertaken show that the TTS pitch is the critical factor in the result quality [7].

A method for prosodic recognition is using Prosodic Markup Language .It allows the synthesizer to determine the intensity of the particular word in the text following tags .For prosodic processing text should be marked with tags XML. Tags indicate to all prosodic attribute values [8].

The different attributes in 'prosody' element like 'rate', 'pitch' and 'contour' are used as specifications to modify predicted phone durations and pitch contour before passing them to the synthesizer[9].

An example of prosodic tagged attributes in Prosodic Mark-up Language is as below:

```
<? Xml version="1.0" encoding="ISO-8859-1"?>
<speak version="1.0" xml: Lang="en-US"
xmlns="http://www.w3.org/2001/10/synthesis">
The price of XYZ is <prosody rate="-10%">
<say-as type="currency">$45</say-as></prosody>
<prosody contour="(0%,+20)(10%,+30%)(40%,+10)">
Good morning
</prosody>
   </speak>
```

In these examples, tags indicate to the two attributes: contour and rate.

MULTI LINGUAL TTS

The first module of the structure can be developed for several languages, the development process directly depends on the language .In some languages, and the transformation is simple because the scripts are orthographic representations of speech sounds. However, in languages like English, the conversion is not straightforward and these languages need large sets of pronunciation rules. In Persian language children's books and some other learning resources, short vowel are marked but generally short vowels do not appear in Persian scripts so a large database includes Persian words with correct pronunciation in Romanization system must be used. Another considerable issue in some languages like Persian and English is Heteronyms .These words have similar spellings but are pronounced differently and have different meaning depending on the context so they need detailed processing. Table 1.shows three examples for Persian heteronyms .If words frequency field is available, with over viewing it, the problem of heteronyms words can somewhat be solved.

Table 1.Persian Heteronyms Samples

Persian script	Latin phoneme	Latin phoneme
ملک	/m eh l k/ means land	/m ae l eh k/ means king
گل	/ g aa l/ means flower	/ g eh l /means mud
ابر	/ ae b r/means cloud	/ ax b aa r/means ultra

The pronunciation of a word in English does not vary according to the sentence it appears. However, in Persian the pronunciation of a word may differ slightly depending on the sentence due to the vowel of the last letter in each word. The vowel of the last letter in a Persian word depends on the function of that word in the sentence. The last letter of a noun can have the vowel /e/ or without voice. The last letter in other words (e.g., verbs, prepositions, conjunctions) is always without vowel. Nouns that govern the genitive case or are qualified by an adjective always have the vowel /e/ as the last letter. Thus, determining the vowel of the last letter requires some grammatical processing [10][11].Table 2 shows an example for this issue in Persian.

Table 2.Noun pronunciation in Persian

Noun	Individual Pronunciation	Sentence	Pronunciation In sentence
Book	Book /b uh k/	He got my book	Book /b uh k/
کتاب(means book)	Ketaab /k eh t ae ae b/	او کتاب من را گرفت He got my book	Ketaabe

Lexical Markup Framework (LMF)

LMF is the International standard for natural language processing (NLP) and machine-readable dictionary (MRD) lexicons. The scope is standardization of principles and methods relating to language resources in the contexts of multilingual communication. LMF contains basic hierarchy of structural skeleton information as a core package for each lexical entry and extensions of the core package, which include morphology, syntax, semantics, multilingual notations, multiword expression patterns, and constraint expression pattern .Using Lexical Markup Framework (LMF) during the word processing to recognize root, postfix, prefix, and morphemes helps to increase speed.

LMF conveys the information which accessibility to them can be solved the ambiguity during text processing such as problem regards the heteronyms words in text.

Implementation OF NON -ENGLISH LANGUAGES TTS using RC8660

The RC8660's integrated TTS processor incorporates RC Systems' DoubleTalk™ TTS technology, which is based on a unique voice concatenation technique using real human voice samples. RC8660 supports Code Page 437 is the character set of the original IBM PC (covers Unites States and Western Europe) and ISO 8859-1/ANSI is the basis for 8-bit character sets (covers Americas, Western Europe, Oceania, and much of Africa Standard Romanization of East-Asian languages) .Both of these character sets are mostly suitable for representing Latin scripts.

Iran System encoding Standard was created by Iran System Corporation for Persian language .This standard could not be used for creating input text for RC8660.Generally enabling RC8660 to speak in non-English languages requires a pronunciation guide for this language, to transcribe the pronunciation rules into exceptional forms.

Since alphabet set other than Latin could not be recognized by RC8660, it is necessary to generate text using Latin characters. Therefore an accurate Romanization or transliteration system must be used .This system provides an unambiguous one-to-one mapping between Latin characters in the UNICODE range and not Latin characters. The Romanization system must be able to preserve both the pronunciation and the written forms of the text. Non-Latin text transliteration requires a reliable system to preserve the orthographic as well as the phonological features of the language .[9].

RC8660 Operating Modes

RC8660 has different modes for operating

- Text Mode(T) '\x01\x54'
- Character Mode(C) '\x01\x43'
- Phoneme Mode(D) '\x01\x44'

Phoneme operating mode disables the text-to-phonetics translator, allowing the RC8660's phonemes to be accessed directly. For example, the word "computer" would be represented phonetically as: k ax m p yy uw dx er.

Phoneme mode can be used to change the stress or emphasis of specific words in a phrase. This is because Phoneme mode allows voice attributes to be modified on phoneme boundaries within each word, whereas Text mode allows changes only at word boundaries.

TTS PREPARATION STEPS FOR LANGUAGES with non-Latin alphabet

To design TTS for the languages with non-Latin alphabet, the size of recognized vocabulary is increased and the speed of the processing is decreased [12].

Moreover, simulating some consonants, which do not have Latin alphabet equivalents, leads to the problems regarding pronunciation particularities.

Before sending non-Latin text to RC8660, the following processes must be undertaken:

- Getting original text as an input
- Breaking text to words by detecting space

- Searching for words in the database for finding correct pronunciation. This database in Persian should be a large lexicon including almost all words with Romanization form of them to detect all short and long vowels in the word [13]. In such this case, text to phoneme conversion is practically dictionary–based.
- Conversion text to Latin script using Romanization system considering correct pronunciation.
- Setting RC8660 in Phoneme mode using command(char D) :

```
echo –en '\x01\x44' > /dev/ttyUSB0
```

Exceptions Dictionaries

The TTS modes of the RC8660 utilize an English lexicon and letter-to-sound rules to convert text speech. Exception dictionaries make it possible to alter the way the RC8660 interprets character strings.

This is useful for correcting mispronounced words and speaking in a non-English language. The pronunciation rules determine which sounds, or phonemes, each character will receive based on its relative position within each word. The integrated Doubletalk text-to-speech engine analyzes text by applying these rules to each word or character, depending on the operating mode in use. Exception dictionaries define exceptions and replace these built in rules. Exception dictionaries can be created and edited with a word processor or text editor that stores documents as standard text (ASCII) files. The dictionary must be compiled into the internal binary format used by the RC8660 before it can be used.

Preparation steps for languages with Latin Alphabet

Before sending non- English Latin based text to RC8660, the following processes must be undertaken:

- Creating Exceptions Dictionaries with file extension dic
- Compiling it into binary format with file extension .dix
- Downloading compiled dictionary to RC8660 using command:

```
echo -en '\x01\247w' > /dev/ttyUSB0
```

This command initializes the RC8660's exception dictionary and stores subsequent output from the host in the RC8660's nonvolatile dictionary memory. The maximum dictionary size is 16 KB.

- Setting RC8660 in Text mode using command (char T)

```
echo –en '\x01\x54' > /dev/ttyUSB0
echo –en '\text\x00'>/dev/ttyUSB0
```

- Enabling Exception Dictionary. The exception dictionary is enabled with this command(char U):

```
echo –en '\x01\x55' > /dev/ttyUSB0
```

If the RC8660 is in Phoneme mode, or if an exception dictionary has not been loaded, the command will have no effect. The exception dictionary can be disabled by issuing one of the mode commands D, T, or C.). The dictionary is disabled by default.

Persian Alphabet

The following tables have been developed using collected data by Council of the Persian language.

Table 3.Persian Vowels

Persian Vowels Type	Latin Equivalents	As in English
Short	A	Hat
Short	E	Men
Short	O	Mode
Long	Aa	Father
Long	I	East
Long	U	Zoo

Table 4. Persian consonants

Persian Consonants	Latin alphabet Equivalents	As in English
پ	P	Pen
ب	B	Boy
م	M	Man
ف	F	Food
و	V	Very
ت ط	T	Time
ض د	D	Day
ص س	S	Sun
ذ ز ظ	Z	Zoo
ر	R	Ray
ن	N	No
ل	L	Love
ج	J	Just
گ	G	Get
ک	K	Key
ی	Y	Yes
ه ح	H	He

Table 5.persian consonants without English equivalent

Persian Consonants	Latin Phonetic Equivalents	As in English
ث	Th	Thin
ش	Sh	She
ژ	Zh	Measure
چ	Ch	Chin
ع		Voiceless glottal plosive
ق غ	Quh	Voiced uvular plosive(as R in French)
خ	Kh	Voiceless uvular fricative (as J in Spanish)

Persian text to speech

Some letters in Persian is a voiceless uvular fricative in English. Thus the pronunciation of a word containing this letter would be ambiguous. To avoid ambiguity in pronunciation, a comprehensive set of letter-to-sound database should be used in the Persian text-to-speech synthesizer.This database is a library of recorded audio files. Each audio file is the sound of a Persian letter. The completed audio library must be stored in the storage of RC8660.The created library file (.sfl) must be compiled (.sfx) and downloaded to the RC8660 through RS232. The Persian TTS applies the letter-to-sound rules on each word of the input text and generates the speech by concatenating audio files.Figure 2 illustrates several processing steps in Implementation of Persian TTS using RC8660 and itshardware system platform.

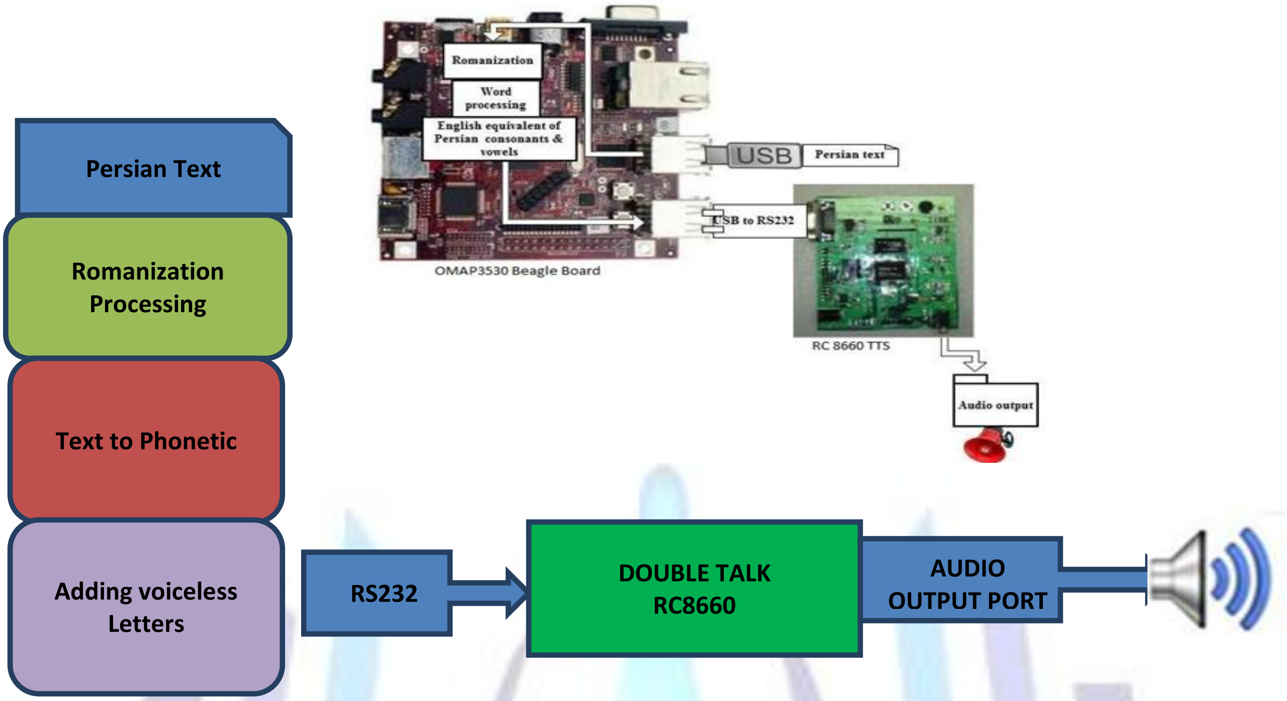

Fig 2.Beagle board (omap3530) as an application processor

RC8660 as an integrated TTS processor

Text To Speech control Playback functions

RC8660 contains two command to stop and start generating speech from text . These command has been used to implement control play back function such as pause, resume, forward and rewind .

Stop command

```
echo -en '\x01\x10' > /dev/ttyUSB0
```

Start command

```
echo -en '\x01\x12' > /dev/ttyUSB0
```

Speech rate calculation of RC8660 has been done by sampling and averaging. Table 6 indicates speech rate for six various samples and average value of them.

Table 6 Speech rate calculation

No of letters	No of words	Speech time	Word /second	Letter/second
3232	493	198	2.48	16.32
1515	217	103	2.1	14.7
1159	189	72	2.6	16.09
294	48	21	2.28	14
526	100	36	2.7	14.61
2697	401	164	2.44	16.44
			Average=2.43	Average=15.36

Control play back functions can be implemented as following:

- Pause: by sending stop command when pause key pressed. Elapsed letters when pause key pressed is calculated by speech rate times elapsed time.
- Resume: by starting TTS task from elapsed letters when pause key pressed again
- Forward: by sending stop command when forward key pressed and starting TTS from :Elapsed letters

when forward key pressed+10*speech rate (it means 10 seconds later)

- Rewind: by sending stop command when rewind key pressed and starting TTS from: Elapsed letters

when rewind key pressed-10*speech rate (it means 10 seconds before)

Conclusion

The research undertaken could make dictionary-based Persian TTS. TTS for each specific language can be made by this method. For achieving better speech quality for various languages TTS, prosody recognition is the significant factor and Prosody Mark–up language information in is useful tool which support TTS Process to collect prosody information regarding the language is needed.

REFERENCES

[1] *Samuel Thomas, M. Nageshwara Rao, Hema A. Murthy, C. S. Ramalingam*, *Natural Sounding TTS based on Syllable-like Units*, in appear in the proceedings of 14th European Signal Processing Conference, Florence, Italy, Sep 2006.units

[2] Carlson, R., Granstrom, B., &Hunnicutt, S. (1982, May 1982). *A multi-language text-to-speech module.* Paper presented at the Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '82..

[3] D. Jurafsky, J.H. Martin(2000) Speech and Language Processing. Pearson Education

[4] A .W.BLACK, K. A. Lenzo (2003) Building synthetic voices

[5] J .Tom´as, F. Casacuberta. (2006) Statistical phrase-based models for interactive computer-assisted translation.

[6] Titov, Ivan &Mcdonald, Ryan.(2008). 2008.A Joint Model of Text and Aspect Ratings for Sentiment Summarization.

[7] M. Plumpe, S.Meredith. (1998) Which is more important in a concatenative text to speech system pitch, duration, or spectral discontinuity?

[8] M.B.Chandak,Dr.R.V.Dharaskar,Dr.V.M.Thakre(2010)Text-to Speech Synthesis with Prosody feature: Implementation

of Emotion in Speech Output using Forward Parsing.

[9] S.Pammi (2003) Prosody control in HMM-based speech synthesis

[10] Hendessi, F., Ghayoori, A., & Gulliver, T. A. (2005).A speech synthesizer for Persian text using a neural network with a smooth ergodic HMM. *4*(1), 38-52. doi: 10.1145/1066078.1066081

[11] B. Sagot, G.Walther (2009) A Morphological Lexicon for the Persian Language

[12] R.Hoffmann,E.Shpilewsky,B.Lobanove,A.Ronzhin(2004)Development of multi sound and multi languages TTS and STT conversion system.

[13] M. A.Mahdavi, (2012) A Proposed UNICODE-Based Extended Romanization System for Persian Texts.

urther development.