Marathi Text to Speech Conversion using Raspberry-Pi Embedded System

Shirbahadurkar S. D., Shiurkar U. D. and Khadake Yogeshwari

Abstract**—**This paper proposes Marathi TTS on Raspberry-Pi Embedded System. Marathi is regional language of Maharashtra (India). Marathi TTS is need for blind or illiterate persons which knows Marathi language. The fastest and effective way of communication is speak out in different language. Adequate combination of words with appropriate grammar provides clear picture of ideas or thoughts that speaker wants to convey. The system includes Python coding which is done on Raspberry-pi for the generation of speech signal based on the user defined input text. The main feature of Marathi TTS is to generate intelligibility and naturalness in output speech, which means that the output sound that is generated at the end of the process should be easily understood and at the same time sound should be natural. Linear predictive coefficients have been used to synthesize speech signal.

Keyword — Concatenation, Raspberry-pi, Speech generation, Speech synthesis, Text to speech.

I. Introduction

The Text to speech synthesis (TTS) system is a system which converts any given text as an input into the sound as an output. Many researchers have developed number of TTS system for foreign languages but from the previous literature it was realized that less work has been done in TTS for Marathi language. Marathi text to speech conversion speech synthesis comes into picture. Speech synthesis is an artificial or computer generated human speech. A system which is used for text to speech synthesis is called speech synthesizer.

A text-to speech system comprises two parts: a front-end and a back-end. The front-end contains two main tasks. First, it converts raw text containing special symbols, numbers and abbreviations into the equivalent words. This process is often called text normalization, preprocessing, or tokenization. Second task is to ascribe phonetic transcriptions to each word, and divide the text into prosodic units like phrases, clauses, and sentences.

II. Literature Review

Tapas Kumar Patra [1] described about how data base can be reduced using phoneme they also mentioned Matlab command that can be used to implement TTS. Mrs. M. R. Repel [2] described the prosody generated for a text to speech system using PSOLA technique.

Shreekanth.T, [3] explained implementation of TTS using Unicode values for Hindi. H. Segi, R. Takou, N. Seiyama and T. Takagi [4] Paper presents technique for weather broadcasting. Shruti Gupta [5] proposed Hindi Text To Speech System, the paper described implementation TTS for Hindi based on JAVA frameworks.

III. Proposed System

In the system, user can see conversion of text messages into speech. Thus the reading aids for the blind, talking aid for the vocally handicapped and training aids and other commercial applications. Vocally handicapped people may type the text from keypad and it will be processed in ARM microcontroller and voice board. In voice board, blind or vocally handicapped people may feed the Input, so that ARM microcontroller will process and output is heard through speaker.

Our system is able to accept UNICODE of input text and identifies each and every character. This system is user friendly and cost effective with Raspberry Pi. The system accepts the text entered through keyboard, converts it to audio format using a synthesized voice for reading out the text quickly translating books, documents and other materials for daily living, especially away from home or office. The system uses concatenative speech synthesis for speech synthesis.

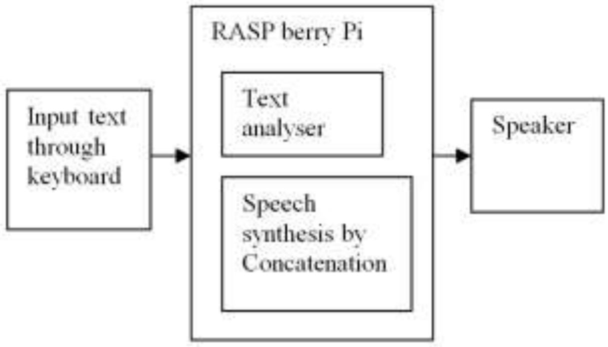
RASP berry Pi
Input text through keyboard
Text analyser
Speech synthesis by Concatenation
Speaker

Fig. 1: Block diagram of Marathi TTS with RASP berry Pi

The system consists of Raspberry-Pi, PC and a speaker. Python language is used for coding the Raspberry-pi. The Raspberry-pi is a small size computer which does not have hard disk as there in PC but has an SD card of 4GB or above for storage purpose. There are numerous interactive application of this small size computer from high quality audio, video playback to playing 3D games etc . It has inbuilt ARM processor in it. ARM processors can be considered as the brain of Raspberry-pi. They are highly efficient in solving complex algorithm in less fraction of time. That is why they are used in small devices like mobile phones, gaming devices and other digital devices. The OS for Raspberry should be installed in SD card apart from this it should store all the program files needed by Raspberry. The SD card should be formatted before installing the OS. The OS should be burnt using Windows 32

Dr. Shirhahadurkar is with Zeal College of Engineering & Research, Pune, India (e-mail: cmanjare@gmail.com, shirsd112@yahoo.in).

Dr. U. D. Shiurkar is with Zeal College of Engineering & Research, Pune, India.

Yogeshwari Khadake is pursuing post graduate degree from Savitribai Phule Pune University, Pune.

https://doi.org/10.17758/EIRAI.IAE1017207

disk imager. The OS in the Raspberry acts as the interface between user and the Raspberry Pi.

There are number of ways through which the input text can be converted into speech like string matching, frequency matching etc. Initially, the GPIO pins should be initialised in order to use them as input or output. This can be coded in python by importing GPIO module. At first the coding is done for matching the alphabets with the user defined alphabets and playing the corresponding audio file. Audio files can be database or voice recordings. Python coding provides one more logic called frequency matching. In this method the frequencies on the basis of repetition of alphabets for each word should be written and then as per the coding logic all the words having matched frequencies should be called and corresponding audio for that word should be played.

Figure 2 shows the flow chart diagram for TTS system.

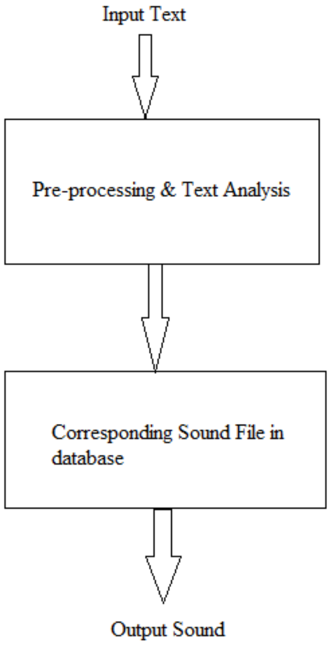
Input Text
Pre-processing & Text Analysis
Corresponding Sound File in database
Output Sound

Fig. 2 Flow chart for TTS operation

IV. SYNTHESIZER TECHNOLOGY

The most important qualities of a speech synthesis system are naturalness and intelligibility. Naturalness describes how closely the output sounds like human speech, and intelligibility is the ease with which the output is understood. The ideal speech synthesizer is both natural and intelligible. Speech synthesis systems usually try to maximize both Characteristics. The primary technology for generating synthetic speech is concatenative synthesis.

A. Concatenative Synthesis

Concatenative synthesis is based on the concatenation of segments of recorded speech. Connecting pre-recorded natural utterances is probably the easiest way to produce intelligible and natural sounding synthetic speech. However, concatenative synthesizers are usually limited to one speaker and one voice and usually require more memory capacity. One of the most important aspects in concatenative synthesis is to find correct unit length.

The selection is usually a trade-off between longer and shorter units. With longer units high naturalness, less concatenation points are achieved, but the amount of required units and memory is increased. With shorter units, less memory is needed, but the sample collecting and labelling procedures become more difficult and complex. In systems units used are usually words, syllables, phonemes.

B. Syllables

A Marathi TTS system using syllables as basic unit of concatenation is presented. The quality of the synthesized speech is reasonably natural. The proposed approach minimizes the co-articulation effects and prosodic mismatch between two adjacent units. Selection of a unit is based on the presiding and succeeding context of the syllables and the position of the syllable. The system implements a matching function which assigns weight to the syllable based on its nature and the syllable with maximum weight is selected as output speech units. We have observed the Syllable is a part of a word that contains a single vowel sound and that is pronounced as a unit.
For example “book” has one syllable, and “reading”
has two syllables. The word “syllable” itself has three
syllables ‘syl-la-ble’.

There are 44 pronunciation sounds (Phonemes) of which 22 sounds are of vowels, and 22 are of consonants. The database of all the pronunciation sounds is created. Once the text is read, for every syllable the corresponding wave files are concatenated and played.
Example:
Word: mahabubnagar
Syllables: ma-ha-b-u-b-na-ga-r

When the above word is played it looks like Stammering and it becomes difficult to make out what is played. This is because every syllable is played separately and it may not be in continuation with the previous sound played. Hence choosing sound unit with proper length is important, so that the word is natural and understandable when synthesized.

LPC analysis and synthesis is used for pitch modification. The analyzer evaluates the LPC for each speech segment equal to the pitch period and in the synthesis mode, the speech samples equal to the samples in one pitch period are reconstructed using LPC inverse synthesis. Instead of synthesizing the same number of speech samples, the synthesizer may synthesize less number of speech samples per segment for speed up and more number of samples per segment for slow down.

It is possible to implement pitch level, pitch range, and pitch contour modification by using a proper modification algorithm in the decoder which inputs pitch values to the synthesizer.

Pitch modification can be implemented on similar lines using any analysis and synthesis techniques which extract the pitch period, such as, channel vocoder etc. The synthesis filter determines the short-term spectral envelope of the synthesized speech and is characterized by the linear prediction (LP) coefficients obtained from LP analysis on the input speech. These coefficients are commonly called LPC coefficients, which may refer generically to any of several different but equivalent parameter sets that specify the synthesis filter [14].

Analysis- Synthesis Technique used for Duration modification:

The speech rate can be modified for the whole phrase. The rate change can be executed by changing the duration of the phonemes. If the duration in consequence of the length reduction is shorter than the frame rate, the phoneme gets dropped.

Spectrum Modification:

Spectrum modification may be done using homomorphic coder and using sinusoidal coder. There is a possibility of male-to-female voice transformation if the fundamental frequency and vocal tract spectrum both are modified.

To establish the level of human performance as a baseline, we first measure the ability of listeners to discriminate between original speech utterances under three conditions: normal, fundamental frequency and duration normalized, and LPC coded. Additionally, the spectral parameter conversion function is tested in isolation by listening to source, target, and converted speakers as LPC coded speech. The results show that the speaker identity of speech whose LPC spectrum has been converted can be recognized as the target speaker with the same level of performance as discriminating between LPC coded speech. However, the level of discrimination of converted utterances produced by the full VC system is significantly below that of speaker discrimination of natural speech [14].

V. CREATION OF DATABASE

There are different factors to be considered While designing a Marathi TTS system that will produce plain speech. The first crucial step in the design of any TTS system is to select the most appropriate units or segments of speech that result in smooth sound. Building the unit record consists of three main phases. First, the natural speech must be recorded so that all used units (phonemes) within all possible contexts (allophones) are included. After this, the units must be labeled from spoken speech data, and finally, the most appropriate units must be chosen. Gathering the samples from natural speech is usually very time-consuming. The implementation of rules to select correct samples for concatenation must also be done very carefully. The voice which is recorded manually contains some delay. This causes a greater time lapse between two consecutives utterances. This makes the speech a bit disagreeable and not natural to listen. Hence there is a need to remove this delay.

In the system we use the Raspbian is the recommended operating system for normal use on a Raspberry Pi. Raspbian is a free operating system based on Debian, optimized for the Raspberry Pi hardware. Raspbian comes with over 35,000 packages: precompiled software bundle in a nice format for easy installation on the Raspberry Pi. Raspbian is a community project under active development, with an emphasis on improving the stability and performance of as many Debian packages as possible. Python is a wonderful and powerful programming language that's easy to use (easy to read and write) and with Raspberry Pi.

VI. RESULTS AND CONCLUSION

In this section we will present analysis of synthesized speech with performance metrics such as Correlation (CORR) and Root Mean Square Error (RMSE) with below significance.

Correlation: This metrics represents the similarity between source voice and modified voice using cross correlation. Therefore, for efficient method it is required that less the correlation more the better method. RMSE: Root Mean Square Error is exactly opposite to Correlation metrics. This represents the difference between original voice and modified voice. We have represented 10 input neutral speech signals and based on observations below table 1 showing results of RMSE and CORR for proposed accent/phrase method.

TABLE 1
RESULTS VERIFICATION

File_Name	Accent/Phrase	
	CORR	RMSE
a.wav	0.3387	14.0029
b.wav	0.4545	14.0055
c.wav	0.4836	14.0033
d.wav	0.4415	14.0003
e.wav	0.2671	14.0049
f.wav	0.5449	14.0006
g.wav	0.4190	14.0046
h.wav	0.4252	14.0015
i.wav	0.6273	14.0016
j.wav	0.5341	14.0020

The proposed Text to speech (TTS) system was designed in order to produce an equivalent acoustic signal which goes in synchronization with the text which is provided as an input to the Raspberry-Pi system. The programming strategy involved the GPIO lines, the database connectivity and the audio amplifiers. This experimental validation can be extended to conjoined sounds and words of higher complexity also. The scope of the Raspberry-Pi in speech synthesis and feature extraction is also quite boundless

REFERENCES

[1] Mr.S.D.Shirbahadurkar, “Marathi Language Speech Synthesizer Using Concatenative Synthesis Strategy (Spoken in Maharashtra, India)”, Second IEEE International sConference on Machine Vision 2009.

[2] Mrs. Madhavi R. Repe, “Natural Prosody Generation in TTS for Marathi Speech Signal”, IEEE International Conference on Signal Acquisition and Processing 2010.

[3] Mrs. Madhavi R. Repe,” Prosody Model for Marathi Language TTS Synthesis with Unit Search and Selection Speech Database”, IEEE International Conference on Recent Trends in information, Telecommunication and Computing 2010

[4] H. Segi, R. Takou, N. Seiyama and T. Takagi, “An automatic broadcast system for a weather report radio program”, IEEE Trans. on broadcasting, vol. 59, no 3, September 2013.

[5] k. Lakshmi Mr. T. Chandra sekhar rao “Design And Implementation Of Text To Speech Conversion Using Raspberry PI” (IJITR) international journal of innovative technology and research Volume No.4, Issue No.6, October – November 2016, 4564-4567.

[6] P.V.N. Reddy “Text to Speech Conversion Using Raspberry-Pi for Embedded System” International Journal of Innovative Research in Science, Engineering and Technology. Vol. 1, Issue 1, November 2012

[7] Mohd Bilal Ganai Er jyoti Arora “Implementation of Text to Speech Conversion Technique” International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 3, Issue 9, September 2015.

[8] Darshna Badhe, P. M. Ghate “Marathi Text to Speech Synthesis – Using Matlab” IJCSN Volume 4, Issue 4, August 2015 ISSN (Online) : 2277-5420.

[9] Sangramsing N. Kayte1, Monica Mundada1,Dr. Charansing N. Kayte Dr. Bharti Gawali “Approach To Build A Marathi Text-To-Speech System Using Concatenative Synthesis Method With The Syllable” Sangramsing Kayte et al. Int. Journal of Engineering Research and Application ISSN: 2248-9622, Vol. 5, Issue 11, (Part - 4) November 2015, pp.93-97.

[10] G. D. Ramteke1 and R. J. Ramteke1 “Text-To-Speech Synthesizer for English, Hindi andMarathi Spoken Signals” British Journal of Applied Science & Technology 15(3): 1-16, 2016, Article no.BJAST.24869 ISSN: 2231-0843, NLM ID: 101664541.

[11] N.sweth a k..anuradha “text-to-speech conversion” International Journal of Advanced Trends in Computer Science and Engineering, Vol.2 , No.6, Pages : 269-278 (2013).

[12] Anand Arokia Raj , Tanuja Sarkar , Satish Chandra Pammi , Santhosh Yuvaraj , Mohit Bansal , Kishore Prahallad, AlanW Black "Text Processing for Text-to-SpeechSystems in Indian Languages" 6th ISCA Workshop on Speech Synthesis, Bonn, Germany, August 22-24, 2007

[13] Hay Mar Htun, Theingi Zin, Hla Myo Tun "Text To Speech Conversion Using Different Speech Synthesis" International journal of scientific & technology research volume 4, issue 07, july 2015.

[14] Dr. Shaila D. Apte, "Speech and Audio Processing ", Wiley Precise Text Book

Dr. Shirhahadurkar received B.E.(EC), M.E.(EC), Ph.D.(EC) degree from Marathwada University, Dr. B.A.M. University, Aurangabad, DOEACC Aurangabad, in 1991,1998, 2010 respectively. He authored 12 International Journals, 14 International Conferences, 3 books. His area of Interest is Speech processing.

Dr. U. D. Shiurkar received B.E.(Instrumentation), M.E.(Instrumentation) and Ph.D. degree from SGGS college of engineering, Nanded in 1985, 1994 and 2009 respectively. His area of interest is signal processing, instrumentation.

Yogeshwari Khadake received B. E. (Electronics and Telecommunication Engineering) from Savitribai Phule Pune University, Pune in 2016. Presently she is pursuing post graduate degree from Savitribai Phule Pune University, Pune.

https://doi.org/10.17758/EIRAI.IAE1017207