Telugu Text to Speech System for Mobile based Systems

Dr. Y. Padma Sai¹, Safia Shaik² and V. Priyanka Brahmaiah³
¹ Professor and Head, ECE, Email: ypadmasai@gmail.com
² M.Tech Student, Email: shaik.safia28@gmail.com
³ Assistant Professor, ECE, Email: priyanka.veeramosu@gmail.com
VNR Vignana Jyothi Institute of Engineering and Technology, ECE Department, Hyderabad, India

Abstract — Speech is the important mode of communication and is the current research topic. The concentration is mostly focused on synthesis and analyzing part. Apart from synthesizing, text to speech system is developed. Speech synthesis is an artificial production of human speech. A text to speech system (TTS) is to convert an arbitrary text into speech. In India, different languages have been spoken each being the mother tongue of tens of millions of people. In this paper, the text to speech system is primarily developed for Telugu, a Dravidian language predominantly spoken in Indian state of Andhra Pradesh. The important qualities expected from this system are naturalness and intelligibility. Telugu TTS can be developed using other synthesis methods like articulatory synthesis, formant synthesis and concatenative synthesis. This paper describes a development of a Telugu text to speech system using concatenative synthesis method on mobile based system OMAP 3530 (ARM Cortex A-8 core) in Linux.

Index Terms — Telugu TTS, Diphone, Prosodic, Phrasing, Lexicon, Concatenative synthesis.

I. INTRODUCTION

Text-to-Speech (TTS) technology deals with the production of synthetic voice output using textual information, thus synthesis technology has become the dominant approach for building naturally sounding text-to-speech systems. In most cases the speech units are phonemes or diphones. The drastic improvement in quality of synthetic speech, namely naturalness and intelligibility, over the years has led to the adoption of TTS as a mainstream technology. As a result, TTS technology is now employed in a wide range of applications, spanning from assistive technology and education, to telecommunications and entertainment. For example, application areas such as assistive aids and tools, speech-to-speech translation, robotics, mobile phones, household devices, navigation and personal guidance gadgets, can largely benefit from the more natural and intuitive means of human computer interaction.

In order for TTS technology to be widely adopted, near-natural voice output quality has to be achieved. Over the last years, significant research progress in the field has contributed towards this goal. Considerable amount of work has been done in conversion of text to speech for languages like English, Japanese, Russian but not much work has been done in TTS for Indian languages especially for Telugu [1]. In order to address the challenge of developing a high quality Telugu TTS system for embedded devices, concatenative synthesis approach has been considered. The function of text-to-speech (TTS) system is to convert an arbitrary telugu

DOI: 02.ITC.2014.5.551
© Association of Computer Electronics and Electrical Engineers, 2014
text to a spoken waveform. Telugu TTS in mainly used for illiterate and it serves as an aid to visually impaired and Language Education. It can also be used in some other applications like talking books and toys, Games, Telecommunication and multimedia etc., Synthesized speech can be produced by different methods. These are classified into three groups [2]. Articulatory Synthesis, which attempts to model the human speech production system directly through articulators like tongue jaw etc., Formant Synthesis, which is done by exciting a set of resonators by voicing sources or noise generator to achieve the desired speech spectrum. Concatenative Synthesis, which uses different pre-recorded samples derived from natural speech. Most of the synthesis systems use formant and concatenative methods. The articulatory method is too difficult for high quality implementations [3], but may arise as a potential method in future. In this work Telugu text to speech system has been implemented using concatenative synthesis for natural sounding telugu speech.

II. CONCATENATIVE SYNTHESIS

Naturalness of synthetic speech produced by state-of-the-art speech synthesis systems is mainly attributed to the use of concatenative speech synthesis that uses phonemes, diphones, syllables, words or sentences as basic speech units. Text is synthesized by selecting appropriate units from a speech database and concatenating them. The concatenation of segments of recorded speech is known as Concatenative synthesis. Connecting pre-recorded natural utterances is the easiest way to produce intelligible and natural sounding speech.

Concatenative synthesis is classified into three main sub-types.

A. Unit selection synthesis
In unit selection synthesis large databases of recorded speech are used.

B. Domain-specific synthesis
Domain specific synthesis concatenates the pre-recorded words and phrases to create complete utterances. It is used in applications like transit schedule announcements or weather reports, railway stations where most of the text remains same and the output is limited to a specific domain.

C. Diphone synthesis
Diphone synthesis considers only diphones occurring in a language and maintains a minimal speech database. In diphone synthesis, only one example of each diphone is contained in the database. The quality of the resulting speech is high and natural [4]. In the present work, diphone synthesis has been adopted to develop telugu TTS.

III. FRAMEWORK OF TELUGU TEXT TO SPEECH SYSTEM

Telugu language is now one of the 5 classical languages of India. Telugu language ranks third by the number of native speakers in India. The block diagram of Telugu Text To Speech (TTS) system is shown in Fig.1. The explanation of each block is as follows.

A. Telugu Text Input
Telugu text to speech system accepts input as Telugu Unicode text [5] (in UTF-8 encoding) and speaks out the text.

B. Text Analysis
Text analysis is nothing but text normalization [6]. This converts raw text into the equivalent of written-out words & isolates the words present in the text. Text normalization then searches for numbers, times, dates, and other symbolic representations. These are analysed and converted into words. Text analysis includes tokenization, token identification and token to word conversion.

1. Tokenization: In this process, it converts the string of characters into a list of tokens. This means that the original text is separated according to the whitespace in between them.
2. Token Identification: Identification of general types of tokens of digits as years, dates, numbers etc.
3. Token to word mapping: This module provides the rules to map the tokens in an utterance to Telugu words. The database contains some default variable telugu dotted abbreviation list.


467
The text pre-processing flow is explained with example “2 తెలుగు” and is shown in Fig.2.

C. Pronunciation Generation

The Pronunciation generation module generates the sequence of basic units using a lexicon of units and letter-to-sound rules.

1. Lexicon:
   It is a subsystem that provides pronunciations for words. It is a list of all speech units like monosyllables, bi syllables and tri syllables. Lexical entries consist of three basic elements. They are a head word, a part of speech and a pronunciation. This entry has internal format, identifying syllable structure, stress markings and phones. Some Examples of lexical entries are shown in the below Table.I.

   ![Table I: Examples of Lexical Entries](image)

2. Letter to Sound Rules: It is practically impossible to assign pronunciation and list all words in a lexicon. The basic letter to sound rule is very simple but powerful enough to build reasonably complex letter to sound. The basic form of a rule is as follows [6]:

   \[(\text{LEFT CONTEXT}) \{\text{ITEMS}\} \text{RIGHT CONTEXT} = \text{NEWITEMS}\]\n
   \(# [c h] C = k\); # - a word boundary, C - the set of all consonants.

   Eg: 1.christmas - #[ch][r] = k, 2.champion - #[ch][a] = ch.

   In these examples ch followed by a consonant is pronounced as ‘K’ and ch followed a vowel is pronounced as ‘Cha’.

D. Prosodic Phrasing

In natural speech, humans tend to group words together with noticeable breaks or disjunctions between them. These groups can be identified as prosodic phrases[7]. Prosodic phrasing plays an important role in
structuring utterances by dividing them into meaningful chunks of information. Text-to-Speech systems should be able to identify these prosodic phrases to produce intelligible and natural sounding speech. In highly inflective languages like Telugu, most words in running texts occur in inflected forms. In an effort to identify linguistically meaningful features that affect prosodic phrasing, a new feature, namely morpheme tag, is defined for telugu language. Morpheme is a meaningful linguistic unit consisting of a word or word element which cannot be divided into smaller meaningful parts. A set of 19 ‘morpheme tags’ are identified that occur at word boundaries (word endings) are shown in the Table II.

TABLE II: LIST OF MORPHEME TAGS IN TELUGU

<table>
<thead>
<tr>
<th>Telugu Morpheme Name</th>
<th>Example word</th>
</tr>
</thead>
<tbody>
<tr>
<td>IO</td>
<td>DhEamIO</td>
</tr>
<tr>
<td>ThO</td>
<td>PattudhalatthO</td>
</tr>
<tr>
<td>Aru</td>
<td>AmAru</td>
</tr>
<tr>
<td>Ndi</td>
<td>Cheppindhi</td>
</tr>
<tr>
<td>Ani</td>
<td>ChEylAlani</td>
</tr>
<tr>
<td>Lu</td>
<td>VisEshAlu</td>
</tr>
<tr>
<td>Nii</td>
<td>PrabhuitvAnni</td>
</tr>
<tr>
<td>Nna</td>
<td>ChErakuunna</td>
</tr>
<tr>
<td>Oni</td>
<td>RAshtraunOni</td>
</tr>
<tr>
<td>Chi</td>
<td>mu.nchi</td>
</tr>
<tr>
<td>Na</td>
<td>Jarigina</td>
</tr>
<tr>
<td>Ki</td>
<td>AdhupalOki</td>
</tr>
<tr>
<td>Ini</td>
<td>PurOgathini</td>
</tr>
<tr>
<td>Ga</td>
<td>Sandharbh.ngA</td>
</tr>
<tr>
<td>Ku</td>
<td>PrAnthAlaku</td>
</tr>
<tr>
<td>Nu</td>
<td>LakshyAlanu</td>
</tr>
<tr>
<td>Pai</td>
<td>Charyapai</td>
</tr>
<tr>
<td>La</td>
<td>Charyala</td>
</tr>
<tr>
<td>.a</td>
<td>Prabhuitv.a</td>
</tr>
</tbody>
</table>

E. Segmental Duration Generation

TTS systems need to generate speech units with appropriate durations in order to produce natural sounding synthetic speech. Duration value for each segment of speech is predefined and it can be changed according to the application [8]. The classification and regression tree (CART) based duration models [9] are used for segmental prediction for Telugu. The CART method is used to build the decision tree such that the branches correspond to questions that minimize the impurity of the sub-clusters.

F. Database

Database defines a telugu diphone set by considering phone features like whether it is vowel or consonant, vowel length, vowel height, vowel frontness, lip rounding, consonant type, place of articulation [9].

G. Waveform generation

The waveform generation component takes as input the phonetic and prosodic information generated by the various components described above, and generates the speech output through speakers.

IV. IMPLEMENTATION

We have developed the Telugu text to speech synthesizer on a Mobile device. The mobile device is a beagle board which consists of OMAP3530 processor with mobile operating system Angstrom ported with the programmable environment supporting component implemented in C++ language. The flow chart of telugu text to speech system on mobile device is shown in the Fig.3.

A. Components of Telugu TTS

1. Mobile Based Device: The Mobile based device is a Beagle board, an OMAP3530 platform designed specifically to address the Open Source Community. Use of the OMAP3530 DCBB72 device which is the
720MHZ version of the OMAP3530. There are many features on this board which are useful for Open Embedded Developers. However, this project uses only few of the features. It has been equipped with a minimum set of features to allow the user to experience the power of the OMAP3530[10]. By utilizing standard interfaces, the Beagle Board is highly extensible to add many features and interfaces.

The high level block diagram consists of OMAP3530[10] processor with SVideo, Touch Screen, Stereo In & Out, USB Host, SD MMC, JTAG, LCD, Expansion pins, Reset & User buttons. Beagle board high level diagram is shown in the Figure 4.

2. Software
- Linux on the Beagle Board
- Angstrom (mobile operating system which is Linux distribution)

3. Porting Angstrom OS: Make two partitions on the SD/MMC card into FAT partition (MLO, u-boot, ulmage) and Ext2 partition.

The five (5) boot phases are
- ROM loads x-load (MLO)
- X-load loads u-boot
- U-boot reads commands
- Commands load kernel(uImage)
- Kernel reads root file system.
V. RESULTS AND DISCUSSIONS

The results have been depicted that the Telugu text to speech system is capable of real time operation and is successfully developed on Mobile based device beagle board, OMAP3530. The telugu text in converted to telugu speech is analysed by various stages. The Telugu TTS system flow is shown in the below Fig.5.

To get an English speech, SayText command should be given with the text inserted under inverted colons. The terminal of the beagle board uttering the telugu speech as output is connected to speakers. To get a telugu speech (voice_telugu NSK_diphone) command is given. When this command is given it calls all the telugu diphones within the database. The input is the telugu text which have been saved in vnrtelugu.txt file and the path has been given in the command. The output speech uttered is natural sounding and clear telugu speech.

VI. CONCLUSION

The full process of converting telugu text to speech is analyzed and various methods used for storing sound and generating voice is studied. It also provides the facility to save the speech file of the input text and can also play any of the previously saved audio file. Various intermediate stages namely, text normalization, prosodic phrasing, pronunciation generation and generation segment durations for converting telugu text to speech is analyzed. It follows the method of diphone concatenation and has a male voice database with diphone as the storage unit. With a natural and clear sounding telugu speech telugu text to speech system have been successfully developed on Mobile device beagle board OMAP3530 which will be useful as assistive tool for visually impaired, illiterate and can be used in many other applications.

Developing text to speech systems for other Indian languages by adding prosody and handling multilingual text Eg: “www.eenadupratibha.net, www.bscacademy.com వంది చందు ఫుటో బ్యాండు చందు ఆయన చందు” is our future work. A Web based application can also be designed which can convert text in any Indian languages into speech.

ACKNOWLEDGEMENTS

The authors acknowledge with thanks to Dr. C.D. Naidu, Principal and Management of VNR VJIET for their constant technical and financial support and encouragement. The research work of developing TTS for Telugu language is part of ITRA.
REFERENCES


[10] Instruction manual provided by the TI Vendor - OMAP3530 Applications Processor by Texas Instruments in 2010.

AUTHORS

Dr. Y. Padma Sai obtained her B.Tech from Nagarjuna University, Guntur, M.E in Systems and Signal Processing and Ph.D in Electronics and Communication Engineering, from Osmania University, Hyderabad. She started her career as a Quality Control Engineer and served for 5 years in M/S. Suchitra Electronics Pvt. Ltd, Hyderabad. Later joined as Lecturer in the Department of ECE in Deccan College of Engineering and Tech, Hyderabad served for one year. She then started working in the Department of ECE VNRVJET on July 1999 and held various positions. Presently, She is the Head of the Department. She has presented 23 research papers in National and International Conferences / Journals. Her areas of research interest are Bio-Medical, Signal and Image Processing. She has received grants from AICTE, DIT and DST to carry out research activities in the department. She is a member of IEEE, ISTE, ISQ and Fellow of IETE. She is executive member of ISTE A.P Section. Her main objective is to impart quality education and learn new technologies and the scope is to fill the gap between Industry and Academics.

Safia Shaik received the B.E degree in electronics and communication engineering from Deccan College of Engineering & Technology, affiliated Osmania University Hyderabad, AP, India, in 2011. She is pursuing the M.Tech in Embedded Systems at VNR Vignana Jyothi Institute of Engineering & Technology, Botapally, Hyderabad, India. Her research interests include Signal Processing and Embedded Systems.

V. Priyanka Brahmainah obtained her B.Tech. Degree from JNT University, Hyderabad in 2007, and M.Tech. in VLSI System design from JNT University, Hyderabad in 2010. She has started her career as Assistant professor and served for 2 years in MLR Institute of Technology & Management, Dundigal, Hyderabad from June 2010 to June 2012. Assistant Professor in the Department of ECE in Gokaraju Rangaraju Institute of Engineering & Technology from July 2012 to November 2012. Assistant Professor in the Department of ECE in VNR Vignana Jyothi Institute of Engineering Technology from December 2012 to till date. She is a Life member of ISTE and IETE. She presented four research papers in International Journals. Her areas of research interest are Bio-Medical, Signal and Image Processing, Human Computer Interface.