Speech Synthesis System for Telugu Language

G. Swathi [1]
M.Tech (SE) Student
VNR VJIET
JNTU University,
Kukatpally,
Hyd , AP, India- 500090

C. Kiran Mai [2]
Computer Science Dept.
VNR VJIET
JNTU University,
Kukatpally,
Hyd, AP, India- 500090

B. Raveendra Babu
Computer Science Dept.
VNR VJIET
JNTU University,
Kukatpally,
Hyd, AP, India- 500090

ABSTRACT

A system which takes input as a sequence of words and converts them to speech. Vowels and consonants are most important in Telugu language. The voices are sampled from real recorded speech. The speech synthesis is handheld by computers and mobile phones.

To build a natural sounding speech synthesis system, it is essential that text processing component produce an appropriate sequence of phonemic units. Generation of sequence of phonetic units for a given standard word is referred to as letter to phoneme rule or text to phoneme rule. The complexity of these rules and their derivation depends upon the nature of the language. In Telugu TTS the input is Telugu text in Unicode.

Speech synthesis is the technique of converting given input text to synthetic speech. Speech synthesis can be used to read written text as in e-mail, SMS, newspapers and can be used by blinds people. Speech synthesis has been widely researched in last four decades. The quality and intelligibility of the synthetic speech produced using the latest methods have been remarkably well for most of the applications.

This project focuses primarily on the process of creating a voice for a concatenative Text-To-Speech system, or altering the TTS systems own standard output voice to sound more like the target voice.

General Terms

Vowels, consonants, phonetic units, Unicode, Grapheme

Keywords

Text processing, speech generation, phoneme, Speech synthesis

1. INTRODUCTION

A system which takes input as a sequence of words and converts them to speech. The aim is to gradually bring the student through basic acoustics, spectrum analysis, vowel and consonant acoustics [1].Vowels and consonants are most important in Telugu language. The voices are sampled from real recorded speech. The goal of Text-to-Speech (TTS) synthesis is to convert arbitrary input text to intelligible and natural sounding [23] speech so as to transmit information from a machine to a person [2].Every language has a different phonetic alphabet and a different set of possible Phonemes and their combinations. A set of phonemes can be defined as the minimum number of symbols needed to describe every possible word in a language. The phonetic alphabet is usually divided in two main categories, vowels and consonants. Vowels are always voiced sounds and they are produced with the vocal cords in vibration, while consonants may be either voiced or unvoiced. Vowels have considerably higher amplitude than consonants and they are also more stable and easier to analyse and describe acoustically. Because consonants involve very rapid they are more difficult to synthesize properly [5]. The speech synthesis is handheld by computers and mobile phones.

The objective of a text to speech system is to convert an arbitrary text into its corresponding spoken waveform [15][18]. Text processing and speech generation are two main components of a text to speech system. We use concatenative based approach to synthesis desired speech through pre-recorded speech waveforms [3] [4].To builds a natural sounding speech synthesis system, it is essential that text processing component produce an appropriate sequence of phonemic units.

Generation of sequence of phonetic units for a given standard word is referred to as letter to phoneme rule or text to phoneme rule. The complexity of these rules and their derivation depends upon the nature of the language. Voiced sounds were simulated with a computer model of the vocal fold composed of a single mass vibrating both parallel and perpendicular to the airflow [6]. In our Telugu TTS the input is Telugu text in Unicode [20].

The work is divided into 3 main modules.

1. Converting Telugu script to Unicode
2. Differentiating Grapheme
   i. Combination of Consonant-vowel
   ii. Combination of Consonant-consonant
3. Generating voice
   i. Identify the Grapheme recognizer
   ii. Identify the Telugu Audio source

Converting Telugu Script to Unicode

In English ASCII characters are used where as In Telugu Unicode characters are used. ASCII takes 8-bits for each character. Unicode takes 16-bits for each character.

Unicode provides a unique number for every character, no matter what the platform, no matter what the program, no matter what the language. The Unicode Standard has been adopted by such industry leaders as Apple, HP, IBM, Just Systems, Microsoft, Oracle, SAP, Sun, Sybase, Unisys and many others.

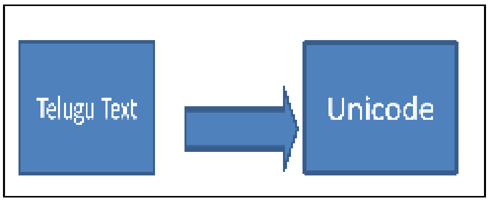


Figure 1. Converting Telugu script to Unicode

Example:

We take an example as a Telugu word మహాభారతం

0c2e -------మ మ

0c39 -------హ

0c3e ------ -ా హా

0c2d -------భ

0c3e -------ా భా

0c30 -------ర ర

0c24 -------త

0c02 -------ం తం

Differentiating Grapheme

In natural speech, durations of phonetic segments are strongly dependent on contextual factors. For synthetic speech to sound natural, the module for computing segmental duration must mimic these contextual effects as closely as possible [7].

Grapheme:

Graphemes are "functional spelling units" encompassing one or more letters of the text input, a grapheme in the text input corresponds to a single phoneme.

Phoneme:

Phones characterize any sound that can be produced by a human vocal tract, if a phone is part of a specific language; it becomes a phoneme of the language. Phonemes are the elementary sounds of a language.

- Generally Telugu script has collection of vowels and consonants. In this project we are going to differentiate them

A character in Indian language scripts is close to syllable and can be typically of the following form: C, V, CV, CCV and CVC, where C is a consonant and V is a vowel. There are about 35 consonants and about 18 vowels in Indian languages [13].

Different Combinations:

- **V--Vowel - అ**
- **C-- Consonant – క**
- **C+V—Consonant+Vowel---** క + ా = కా(కాని)
- **C+C----Consonant+Consonant---మ+మ=మ్మ(అమ్మ)**
- **C+C+V----Consonant+Consonant+Vowel---** **స+ వ+ ా= స్వా(స్వాతి)**
- **C+C+C--Consonant+Consonant+Consonant—** ***ష+ట+ర = ష్ట్ర(రాష్ట్రం)***

Example:

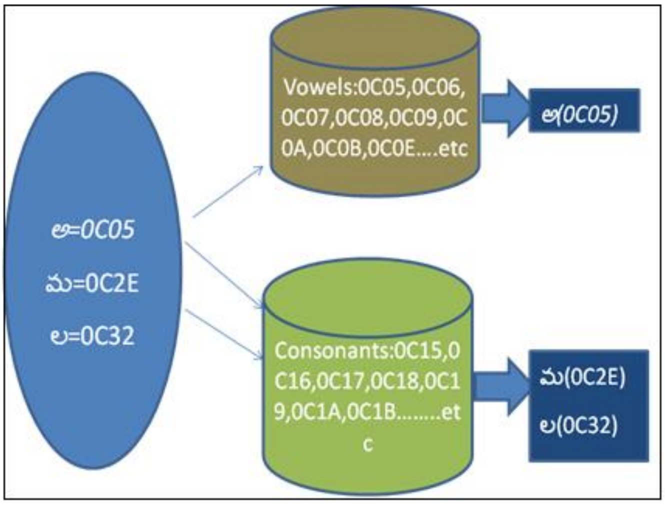


Figure 2. Differentiating grapheme

Generating Voice

The function of Text-To-Speech (TTS) system is to convert the given text to a spoken waveform. This conversion involves text processing and speech generation processes. These processes have connections to linguistic theory, models of speech production, and acoustic-phonetic characterization of language. To build a voice/speech for a language text, the steps involved are as follows

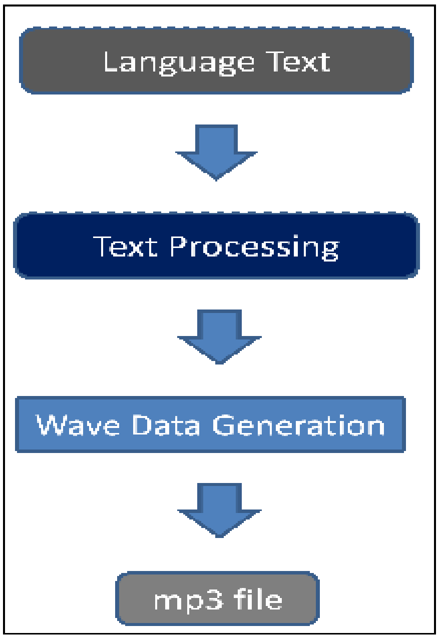


Figure 3. Generating voice

Text processing including end-of-sentence detection, text normalization. Word pronunciation, including the pronunciation of names and the disambiguation of homographs [10].

In this approach, the pre-recorded speech segments which are to be used in the synthesizer are stored exactly as how it is recorded. Additional information of the speech waveform is attached to the sound to provide proper annotation of the speech waveform [12].

2. HISTORY

Speech synthesis is the artificial production of human speech. A computer system used for this purpose is called a speech synthesizer, and can be implemented in software or hardware products. A text-to-speech (TTS) system converts normal language text into speech; other systems render symbolic linguistic representations like phonetic transcriptions into speech.

The overview of the problems that occur during text-to-speech (TTS) conversion and describe the particular solutions to these problems taken within the AT&T Bell Laboratories TTS system. In addition to discussing the linguistic and speech analysis issues that must be addressed in a high-quality TTS system [9].

Synthesized speech can be created by concatenating pieces of recorded speech that are stored in a database. Systems differ in the size of the stored speech units, a system that stores phones or diphones provides the largest output range, but may lack clarity. For specific usage domains, the storage of entire words or sentences allows for high-quality output. Alternatively, a synthesizer can incorporate a model of the vocal tract and other human voice characteristics to create a completely "synthetic" voice output. The quality of a speech synthesizer is judged by its similarity to the human voice and by its ability to be understood. An intelligible text-to-speech program allows people with visual impairments or reading disabilities to listen to written works on a home computer. Many computer operating systems have included speech synthesizers since the early 1990s.

A text-to-speech system is composed of two parts a front-end and a back-end. The front-end has two major tasks. First, it converts raw text containing symbols like numbers and abbreviations into the equivalent of written-out words. This process is often called text normalization, pre-processing, or tokenization. The front-end then assigns phonetic to each word, and divides and marks the text into prosodic units, like phrases, clauses, and sentences. The process of assigning phonetic transcriptions to word is called text-to-phoneme or grapheme-to-phoneme conversion [21].

Phonetic transcriptions and prosody 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. In certain systems, this part includes the computation of the target prosody (pitch contour, phoneme durations), which is then imposed on the output speech.

3. SYSTEM ARCHITECTURE

The main purpose of the project is to convert an arbitrary text into its corresponding spoken waveform. Text processing and speech generation are two main components of a text to speech system. To build a natural sounding speech synthesis system, it is essential that text processing component produce an appropriate sequence of phonemic units. Generation of sequence of phonetic units for a given standard word is referred to as letter to phoneme rule or text to phoneme rule. The complexity of these rules and their derivation depends upon the nature of the language. In Telugu TTS the input is Telugu text in Unicode.

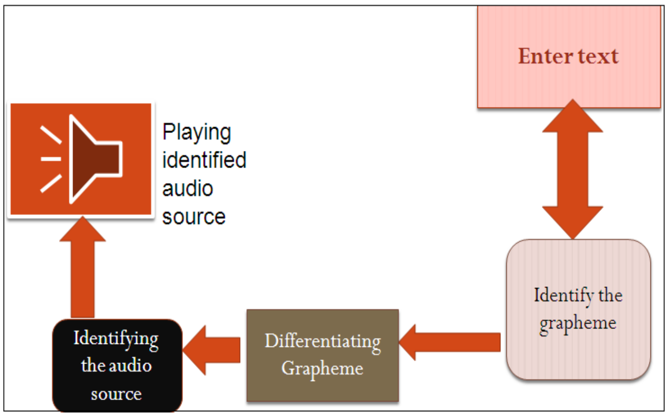


Figure 4: Architecture of the Speech synthesis system

Finite State Machine

In this finite state machine initial stage we take the bellow example in the first we take initial state as a consonant then we check the next letter. The next letter is a vowel it goes to the vowel diacritics state. If the letter is complete it is the finite statement then display the phoneme. Other letters also check the finite state machine similarly.

The use of finite-state models of morphology also makes for easy interfacing between morphological information and finite state models of syntax. One obvious finite-state syntactic model is an n-gram model of part-of-speech sequences [8].

Whereas some phonemes [19] can affect the articulation over several phonemes. Syllables are alternative units for concatenative synthesis [14].

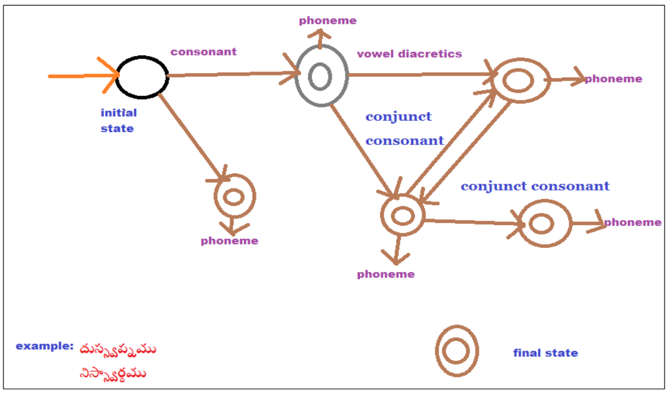


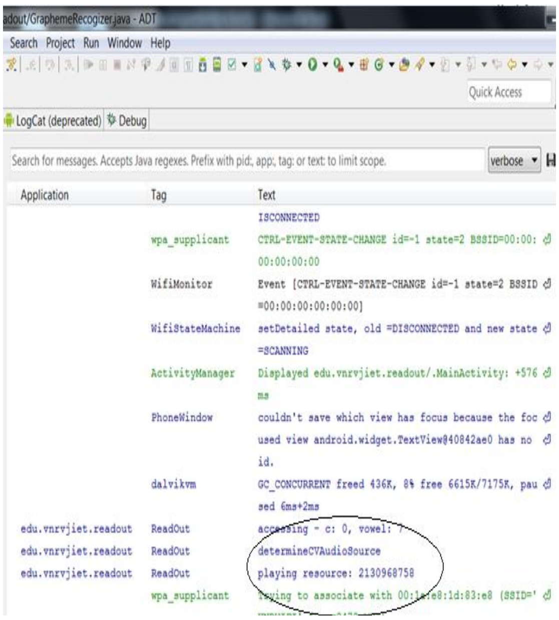
Figure 5: finite state machine of the system

4. RESULTS

Telugu 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 efficiency of this approach for Telugu language and found that the performance of this approach is better.

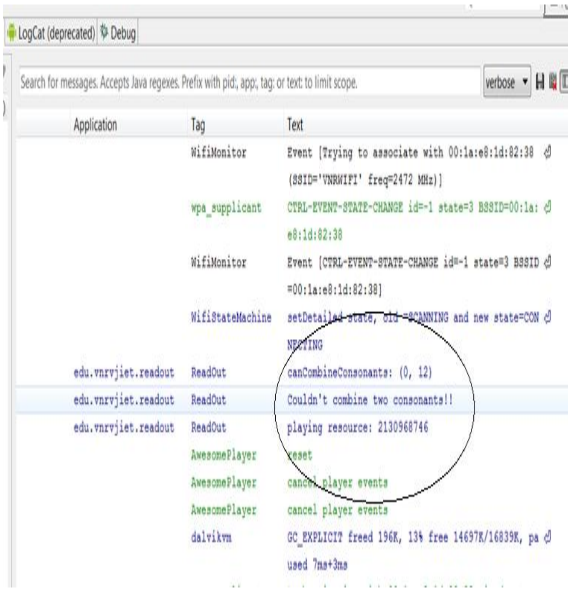
The results showed that there is a strong correlation between the values of the source parameter in the vowel midpoint and the vowel duration. The same parameters tend to decrease on vowel onsets and to increase on vowels offsets. This seems to indicate a prosodic nature of these parameters requiring special treatment in concatenative-based TTS systems that use source modification techniques, such as pitch synchronous overlap add and multipulse [11].

This project focuses primarily on the process of creating a voice for a concatenative Text-To-Speech system, or altering the TTS systems own standard output voice to sound more like the target voice.


adout/GraphemeRecognizer.java - ADT
Search Project Run Window Help
Quick Access
LogCat (deprecated)
Debug
Search for messages. Accepts Java regexes. Prefix with pid:, app:, tag: or text: to limit scope.
verbose

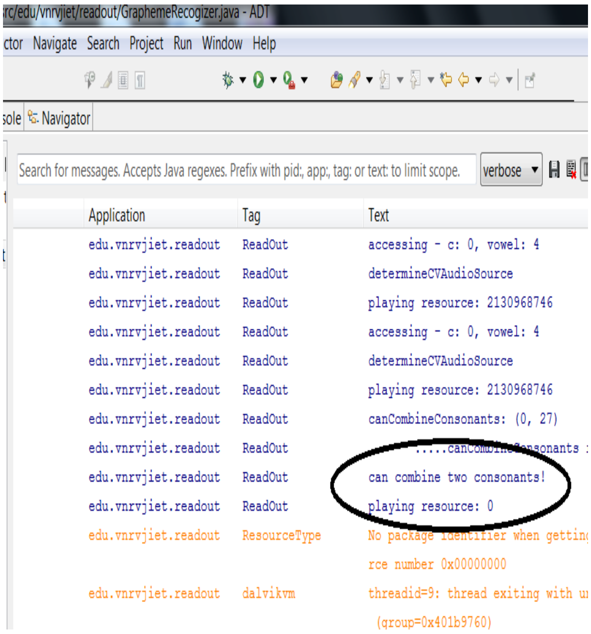
```
Application            Tag                Text
                                          ISCONNECTED
                       wpa_supplicant     CTRL-EVENT-STATE-CHANGE id=-1 state=2 BSSID=00:00:
                                          00:00:00:00
                       WifiMonitor        Event [CTRL-EVENT-STATE-CHANGE id=-1 state=2 BSSID
                                          =00:00:00:00:00:00]
                       WifiStateMachine   setDetailed state, old =DISCONNECTED and new state
                                          =SCANNING
                       ActivityManager    Displayed edu.vnrvjiet.readout/.MainActivity: +576
                                          ms
                       PhoneWindow        couldn't save which view has focus because the foc
                                          used view android.widget.TextView@40842ae0 has no
                                          id.
                       dalvikvm           GC_CONCURRENT freed 436K, 8% free 6615K/7175K, pau
                                          sed 6ms+2ms
edu.vnrvjiet.readout   ReadOut            accessing - c: 0, vowel: 7
edu.vnrvjiet.readout   ReadOut            determineCVAudioSource
edu.vnrvjiet.readout   ReadOut            playing resource: 2130968758
                       wpa_supplicant     Trying to associate with 00:1a:e8:1d:83:e8 (SSID='
```

Figure 6: Identifying the consonant and vowel


LogCat (deprecated)
Debug
Search for messages. Accepts Java regexes. Prefix with pid:, app:, tag: or text: to limit scope.
verbose

```
Application            Tag                Text
                       WifiMonitor        Event [Trying to associate with 00:1a:e8:1d:82:38
                                          (SSID='VNRWIFI' freq=2472 MHz)]
                       wpa_supplicant     CTRL-EVENT-STATE-CHANGE id=-1 state=3 BSSID=00:1a:
                                          e8:1d:82:38
                       WifiMonitor        Event [CTRL-EVENT-STATE-CHANGE id=-1 state=3 BSSID
                                          =00:1a:e8:1d:82:38]
                       WifiStateMachine   setDetailed state, old =SCANNING and new state=CON
                                          NECTING
edu.vnrvjiet.readout   ReadOut            canCombineConsonants: (0, 12)
edu.vnrvjiet.readout   ReadOut            Couldn't combine two consonants!!
edu.vnrvjiet.readout   ReadOut            playing resource: 2130968746
                       AwesomePlayer      reset
                       AwesomePlayer      cancel player events
                       AwesomePlayer      cancel player events
                       dalvikvm           GC_EXPLICIT freed 196K, 13% free 14697K/16839K, pa
                                          used 7ms+3ms
```

Figure 7:can't combine the consonants


rc/edu/vnrvjiet/readout/GraphemeRecognizer.java - ADT
ctor Navigate Search Project Run Window Help
sole
Navigator
Search for messages. Accepts Java regexes. Prefix with pid:, app:, tag: or text: to limit scope.
verbose

```
Application            Tag            Text
edu.vnrvjiet.readout   ReadOut        accessing - c: 0, vowel: 4
edu.vnrvjiet.readout   ReadOut        determineCVAudioSource
edu.vnrvjiet.readout   ReadOut        playing resource: 2130968746
edu.vnrvjiet.readout   ReadOut        accessing - c: 0, vowel: 4
edu.vnrvjiet.readout   ReadOut        determineCVAudioSource
edu.vnrvjiet.readout   ReadOut        playing resource: 2130968746
edu.vnrvjiet.readout   ReadOut        canCombineConsonants: (0, 27)
edu.vnrvjiet.readout   ReadOut            .....canCombineConsonants :
edu.vnrvjiet.readout   ReadOut        can combine two consonants!
edu.vnrvjiet.readout   ReadOut        playing resource: 0
edu.vnrvjiet.readout   ResourceType   No package identifier when gettin
                                      rce number 0x00000000
edu.vnrvjiet.readout   dalvikvm       threadid=9: thread exiting with u
                                        (group=0x401b9760)
```

Figure 8: Identifying the consonant and consonant

Figure 9: Playing the audio source

5. CONCLUSION

Speech synthesis can be used to read written text, SMS, newspapers and can be used by blind people. This project focuses primarily on the creating a voice for a Text-To-Speech system, or the TTS systems own standard output voice to sound more like the target voice. The Speech files for Telugu syllable and words are recorded and stored in mp3 format. The Speech files are created for naturalness of the synthesized output. Telugu TTS system using syllables as basic unit of concatenation [24] is presented. The quality of the synthesized speech is reasonably natural.

Speech synthesis techniques [25], it is much easier to build a voice in a language with fewer sentences and a smaller Speech [17]. 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 efficiency of this approach for Telugu language and found that the performance of this approach is better.

This project focuses primarily on the process of creating a voice for a concatenative Text-To-Speech system, or altering the TTS systems own standard output voice to sound more like the target voice.

6. FUTURE ENHANCEMENTS

Future work will mainly focus on improving the naturalness [25] of the synthesizer [16]. Work is in progress to improve the prosody modules. A speech corpus containing 2 hours of speech has been already recorded. The material is currently being segmented, and labelled. We are also planning to improve the duration model using the data obtained from the annotated speech corpus. A number of other ongoing projects are aimed at developing a POS tag set, POStagger and a tagged corpus for Sinhala. Further work will focus on expanding the pronunciation lexicon. At present, the G2P rules are incapable of providing accurate pronunciation for most compound words. Thus, we are planning to construct a lexicon consisting of compound words along with common high frequency words found in our Sinhala text corpus, which are currently incorrectly phonetized.

The generated speech shows distortion at the concatenation point of two syllables. If this distortion is significant then it would loose the naturalness. In future we give the Telugu text that converted into the voice.

7. ACKNOWLEDGMENT

Firstly, I would like to express my immense gratitude towards my institution **VNRVJIET** which created a great platform to attain profound technical skills in the field of Computer Science, there by fulfilling my most cherished goal.

I am thankful to **Dr. D.N.Rao** (General Secretary), **Dr.C.D.Naidu** (Principal) and **Management** for providing the necessary infrastructure and labs.

I express my sincere gratitude and thankfulness towards my Project Guide, **Dr.C.Kiranmai** (Vice Principal & Professor), **Dr. B.Ravindra Babu** (Professor and Head of the Department, CSE Department,), **Mr. G.Ramesh Chandra** (Head of the Department, RCC, and Associate Professor), **Dr. M.Devi Prasad** (IT Department, and Professor), **Dr.G.Sharada** (Course Coordinator, CSE Department, Associate Professor), **Mrs.R.Vasavi** (Project Coordinator, CSE department, Assistant Professor) of VNRVJIET for extending their co-operation in doing this project.

My special thanks to all the Faculty members of the Department for their valuable advises at every stage.

I would like to express my deep sense and earnest thanks giving to my dear Parents and my family members for their moral support and kind cooperation in doing the project.

I would also like to thank all my friends, whose direct or indirect help has enabled me to complete this project.

8. REFERENCES

[1] C. Bickley, A.Syrdal, and J.Schroeter, ‘‘Speech Synthesis,’’ in The Acoustics of Speech Communication, J.M. Picket, Ed., Boston, NY: Allyn and Bacon, 1998.

[2] T. Dutoit,An Introduction to Text-to-Speech Synthesis, Dordrecht/Boston/London: Kluwer Academic Publishers, 1997.

[3] Lakshmi A, Hema A Murthy. A Syllable Based Continuous Speech Recognizer for Tamil. In Proc. of the 2nd Int. Workshop on East-Asian Language Resources and Evaluation,2009.

[4] Ö. Salor, B. Pellom and M. Demirekler, "Implementation and Evaluation of a Text-to-Speech Synthesis System for Turkish", Proceedings of Eurospeech-Interspeech 2003, Geneva, Switzerland, 2003, pp. 1573-1576.

[5] S. Lemmetty, Review of Speech Synthesis Technology, MSc. thesis, Helsinki University of Technology, 1999.

[6] K. Ishizaka and J.L. Flanagan, ‘‘Synthesis of voiced sounds from a two-mass model of the vocal cords,’’ Bell Syst. Tech. J., vol. 51, no. 6, pp. 133–1268, 1972.

[7] van Santen J.P .H. (1994): " Assignment of seg-mental duration in text-to-speech synthesis". Com-puter Speech and Language 8, 95-128

[8] Sproat R. (1995): " A finite-state architecture for tokenization and grapheme-to-phoneme conver-sion for multilingual text analysis". In F rom text to tags: Issues in multilingual language analysis. Proc. ACL SIGDAT W orkshop (Dublin, Ireland), 65-72

[9] Sproat R., Olive J. (1995): "Text to speech syn-thesis". AT&T T echnical Journal 74(2), 35-44

[10] Sproat R., Olive J. (1996): " A modular architec-ture for multi-lingual text-to-speech". In J. van Santen, R. Sproat, J. Olive and J. Hirschberg (eds.), Progress in speech synthesis (Springer , New Y ork).

[11] T alkin D., Rowley J. (1990): "Pitch-syn-chronous analysis and synthesis for TTS systems".Proc. ESCA W orkshop on Speech Synthesis (Autrans,France), 55-58.

[12] A.M. Zeki and N. Azizah, "A Speech Synthesizer for Malay Language", National Conference on Research and Development in Computer Science, Selangor, Malaysia, October 2001.

[13] S P Kishore, Rohit Kumar and Rajeev Sangal, "A Data Driven Synthesis Approach For Indian Languages using Syllables as BasicUnit", in Proceedings of Intl. Conf. on NLP (ICON) 2002, pp. 311-316, Mumbai, India, 200.

[14] O. Fujimura and J. Lovins, ‘‘Syllables as concatenative phonetic elements,’’ inSyllables and Segments, A.Bell and J.B. Hooper, Eds., New York: North-Holland, 107–120, 1978.

[15] BlackA.W.,ZenH.,andTokudaK.,"Statistical parametric speech synthesis," in Proceeding sofIEEEInt.Conf. Acoust., Speech,and Signal Processing, Honolulu,USA, 2007.

[16] Alan W Black, Paul Taylor, "Automatically Clustering similar units for unit selection in speech synthesis", Proceedings of Eurospeech 97.

[17] ZenH.,NoseT.,YamagishiJ.,SakoS.,MasukoT.,Black A.W., andTokudaK.,"The hmm-based speech synthe sis system version2.0," in Proc.ofISCASSW6, Bonn, Germany,2007.

[18] A.W. Black, and K.A. Lenzo, Building Synthetic Voices, Language Technologies Institute, Carnegie Mellon University and Cepstral LLC.

[19] B. Williams, R.J. Jones and I. Uemlianin, "Tools and Resources for Speech Synthesis Arising from a Welsh TTS Project", Fifth Language Resources and Evaluation Conference (LREC), Genoa, Italy, 2006.

[20] C. Kamisetty and S.M. Adapa, Telugu Festival Text-to-Speech System.

[21] A. Wasala, R. Weerasinghe and K. Gamage, "Sinhala Grapheme-to-Phoneme Conversion and Rules for Schwa epenthesis", Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, Sydney, Australia, 2006, pp. 890-897.

[22] J.B. Disanayaka. 1991. The Structure of Spoken Sinhala, National Institute of Education, Maharagama.

[23] Marian Macchi, Bellcore,"Issues in text-to-speech Synthesis" In Proc. EEE International Joint Symposia on Intelligence and Systems, pp.318-325, 1998.

[24] A. Hunt, & A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database", In Proc. of EEE int. Conference acoust, speech, and signal processing, vol. 1, pp. 373–376, 1996.

[25]Carlson, R., & Nord, L."Vowel dynamics in a text-to-speech system - some considerations". In Proceedings Eurospeech '93 (pp. 1911-1914). Berlin, 1993.

[26]Anupam Basu, Debasish Sen , Shiraj Sen and Soumen Chakraborty "An Indian Language Speech Synthesizer – Techniques and Applications" National Systems Conference, Indian Institute of Technology, Kharagpur, december 17-19, 2003.