Urdu Text to Speech Synthesizer

By

Muhammad Hassan Siddiqui

MSCSF15M005

Supervised by

Dr. Muhammad Kamran Malik

Assistant Professor, PUCIT

(June, 2018)

Punjab University College of Information Technology,

University of the Punjab, Lahore, Pakistan.

Urdu Text to Speech Synthesizer

A THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE

DEGREE OF

MASTER OF PHILOSOPHY

IN

COMPUTER SCIENCE

By

Muhammad Hassan Siddiqui

MSCSF15M005

Supervised by

Dr. Muhammad Kamran Malik

Assistant Professor, PUCIT

(June, 2018)

Punjab University College of Information Technology,

University of the Punjab, Lahore, Pakistan.

Evaluation of M. Phil. Thesis

We have evaluated the M. Phil. thesis titled

Urdu Text to Speech Synthesizer

Submitted by Mr. **Muhammad Hassan Siddiqui**, **MSCSF15M005**, session 2015-2018 in partial fulfillment of the M. Phil. degree in Computer Science. We have also assessed the candidate through viva-voice.

We are satisfied with the thesis and performance of the candidate in the examination and are of the opinion that she fulfills the requirements as set in the rules and regulations for the M.Phil. degree in Computer Science at the University of the Punjab.

Thesis Supervisor: Dr. Muhammad Kamran Malik
Assistant Professor
Punjab University College of Information Technology
University of the Punjab, Lahore

External Examiner: Dr. NAME
Assistant Professor
Department of Computer Science
COMSATS Institute of Information Technology, Lahore

Principal of the College: Dr. Syed Mansoor Sarwar Principal,
Punjab University College of Information Technology
University of the Punjab, Lahore

UNIVERSITY OF THE PUNJAB

Author: **Muhammad Hassan Siddiqui**

Title: **Urdu Text to Speech Synthesizer**

Department: **Punjab University College of Information Technology**

Degree: **M. Phil. (Computer Science)**

Permission is herewith granted to University of the Punjab to circulate and to have copied for non-commercial purposes, at its discretion, the above title, upon the request of individuals or institutions.

Signature of the Author

Dedicated to

Abstract

Text to speech synthesis system is system which takes raw text as input and converts it into speech signal. This is done by concatenation of small speech segments called phonetic strings of words or Statistical parametric speech synthesis which uses parameters to describe speech. In this technique, model is learned from speech data using Hidden Markov model (HMM) or Deep Neural Networks (DNN).

This paper describes development of Festival TTS system based Urdu text to speech system using Hidden Markov model (HMM). It describes Urdu text preprocessor system used to process numbers, dates and time text in input data and how Festvox voice package is generated for Urdu. In the end, evaluation of system is conducted using DRT, MRT and MOS tests to get performance of the system. .

Keywords: Text to Speech, Urdu Text Preprocessor, Hidden Markov model, Festival, Festvox

Acknowledgements

Computational modeling is branch of computer science which deals with multiple disciplines. It assists other domains in understanding complex systems and phenomena by providing theory, tools and technology to model and simulate related systems and phenomena. In complex systems, behavior of an individual can have butterfly effect and can become root cause of an emergent phenomenon. Interaction of drivers with each other and surrounding environment forms the dynamics of traffic flow. Hence global effects of a traffic flow depend upon behavior of a single driver. In this research.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Speech Synthesis

Speech is most important medium of conveying opinions and expressing feeling and thoughts. Human convert their thought into speech by using words, phrases and sentences in order to communicate with each other [1]. Speech is produced when air is exhaled by the lungs and vibrations are produced by air, these vibrations got a proper waveform shape by glottal cords and vocal tract. Text to Speech synthesis is the process of conversion of raw text into speech signals. It works by concatenation of small segments of recorded speech called phonemes [2].

The TTS system comprises of two main stages. One is called Natural language Processing (NLP) and other is called Speech Synthesis (SS). This is shown in figure 1-1.

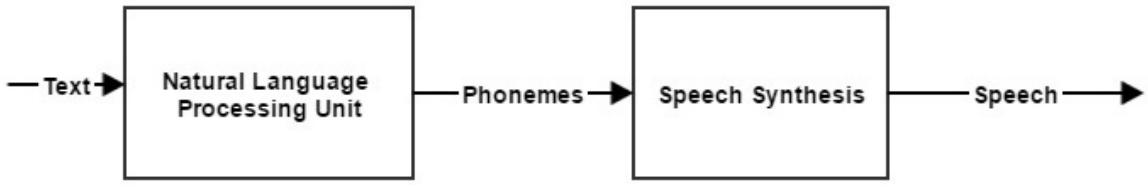

Figure 1-1: TTS Block Diagram

In NLP unit, text is first converted into string of letters and then word boundaries are marked by tokenizer. This is called normalization of text. Normalized data is then

converted into phonetic strings with the help of letter to sound rules after which Syllabifier marks syllable boundaries. Sound change rules are applied on the syllabified date. Language modeling techniques are also applied for finding context in which a specific word is used. As human has tendency to recognize basic rules for his native language, it is easy to judge context of a word in a sentence and what should be correct pronunciation of that word with respect to its context. For example, it can be guessed easily that in a sentence is used for (moment) or (bridge) for any native speaker of Urdu. Last stage of NLP is stress intonation marker which adds stress and intonation to the text. Speech Synthesis unit converts symbolic information received from NLP unit into audible speech with the help of different Digital Signal Processing Techniques. The quality of speech synthesis system is detected by naturalness and intelligibility of the speech.

Partially blinded or fully blinded people usually suffer while using computer technology when there is no assistant or computer is not enough interactive. Due to which text to speech systems are becoming necessity of modern life. These systems increase the degree to which blind people can interact with sighted people [3] and could boost up their hope to survive in this world gracefully [4]. Many applications of speech synthesis are emerging such as machines that read for blinds, aids for handicaps, computers that interact with user through speech. For all these applications a text to speech that convert text to speech are used [5].

In digital world there are some people who can read and understand different languages and some who cant understand languages except their own languages. Speech to text conversion system can also provide a facility to exchange information between people speaking different languages [2]. TTS systems are also needed to reduce the extinction of minority languages. As minority languages of the world are facing challenge of extinction considerable efforts are going on from last few years for their survival. Fon language is spoken in Republic of Benin and some other regions of Africa and it is also facing challenge of extinction [6]. The Xitsonga is spoken in more than three African countries. TTS system of such languages will help lot of people of different literacy level [7]. Urdu is national language of Pakistan and it is spoken by more than 100 million people across the world [8]. A Text-to-Speech (TTS) system for Urdu will be very helping for visually impaired, handicapped and illiterate people.

For human, the task of speech synthesis is not difficult one as they have basic knowledge

of their language but for computer some other method has to be implemented for this task. When we talk about TTS systems speech types and procedure for synthesis, strategies or modules used to develop systems etc. are important to consider. Different types of speech exist such as isolated word (process single word at a time), connected words (isolated words but separated with least gap), continuous speech (permit client to talk while computer is processing content) and spontaneous speech (deals with variety of words that are used rarely) as well as two types of speaker model were presented independent and dependent of clients or speaker specifications. Vocabulary is also characterized according to size such as small vocabulary, medium vocabulary, large vocabulary, very large vocabulary and out-of-vocabulary. Below are the major speech generation techniques.

1.2 Types of Speech Synthesis

For the process of speech synthesis, three types of processes are used.

1.2.1 Formant Synthesis

In Formant Synthesis, simple model of speech production and set of rules are used to generate speech. But it is very difficult to accurately describe the process of speech generation in set of rules

1.2.2 Concatenative Synthesis

Concatenative Synthesis small units are selected from carrier sentences which then join to form speech of complete sentence. These small units are called phonemes. These are the units which collectively describe correct pronunciation of a word. This process is easy as compared to previous one as number of such phonemes is limited for any language. For English, there are 44 such phonemes. Similarly in Urdu, there are 44 consonants, 8 long vowels, 7 long nasal vowels, 3 short vowels and many diphthongs [9]. This reduce distortion but it can decrease the naturalness. Thats why the derived synthetic speech may not resemble the donor speaker in training database [10].

1.2.3 Statistical Parametric Speech Synthesis

Statistical parametric speech synthesis is another approach which uses parameters to describe speech. In this technique, model is learned from speech data. This technique works better than concatenative technique [11].

1.3 Quality

The intelligibility and naturalness is the measure of quality of the synthesized speech [12]. There is lots of experimentation over naturalness of voice as a result of TTS systems. In todays world different segments are recorded and then concatenated for completing a message. A collection of speech words is collected and maintained in database by using a reader who reads large series of text. In these kind of system, to maintain the consistency the speaker speaks in a single style and keep in mind the distance from microphone and other factors to avoid the inconsistency. This type of TTS system is not required at all as the need is to have a system which can be expressive and convey message with proper expressions and styles. Work is performed to build a system that can convey the message according to the needs of the users. A single style of communication can lead towards wrong messages and can cause other problems of understandings. For example, it is not appropriate to convey a good news and bad news in a same style and manner. Similarly, it is not acceptable to ask a question in neutral way of communication [4]. Multiple techniques like linear regression and neural networks were applied to get the results with improvements. Concatenation techniques are applied to get fully expressive and stylish messages for end users. By using concatenation technique users can customize and add styles and expression through provided Speech Synthesis Markup Language (SSML) [4]. Timing of events in speech is also important as timing of events in speech signals are affected by some contextual factors like phone identity factors. These factors make it difficult to control timing of events [13]. There are some approaches which have been proposed to control timing of events like linear regression [14] and tree regression [15]. A new technique is proposed in [13] in which timing of events is controlled by multi-dimensional Gaussian distribution based Hidden Markov model.

1.4 Architecture

Text to speech is a way of communication and transferring information using words and styles of speaking [4]. It has two processes which are text processing and speech generation. In text processing given input text is processed so that to get appropriate chain of phonemic units. Speech generator takes these units as inputs and convert them into synthetic speech by selection of a unit from large corpus TTS system for small database is easier to implement but not in good quality [16] [17] [18]. Different researchers and developers use different strategies to develop TTS system such as in [19], author interpret text to speech system as, it converts raw text into intelligible speech signals by following two sub processes called High-level synthesis and Low-level synthesis. High-level synthesis converts text into phonetic strings and Low-level synthesis converts these strings into speech signals [19]. In [20], TTS system is divided in three modules.

1. Natural language Processing
2. Text Parameterization
3. Speech Synthesis

Natural Language processing unit converts text into phonetic strings. The second and third stages use these phonetic strings and convert them into speech signals. This is shown in figure 1-2.

In [21], TTS system is implemented by following four modules in sequence.

1. Text Analysis
2. Word Pronunciation
3. Phonetic Interpretation
4. Speech Signal Generation

In text analysis, the inputted text is segmented into sentences and later on to words. These words are then categorized according to their syntactic and contextual meaning.

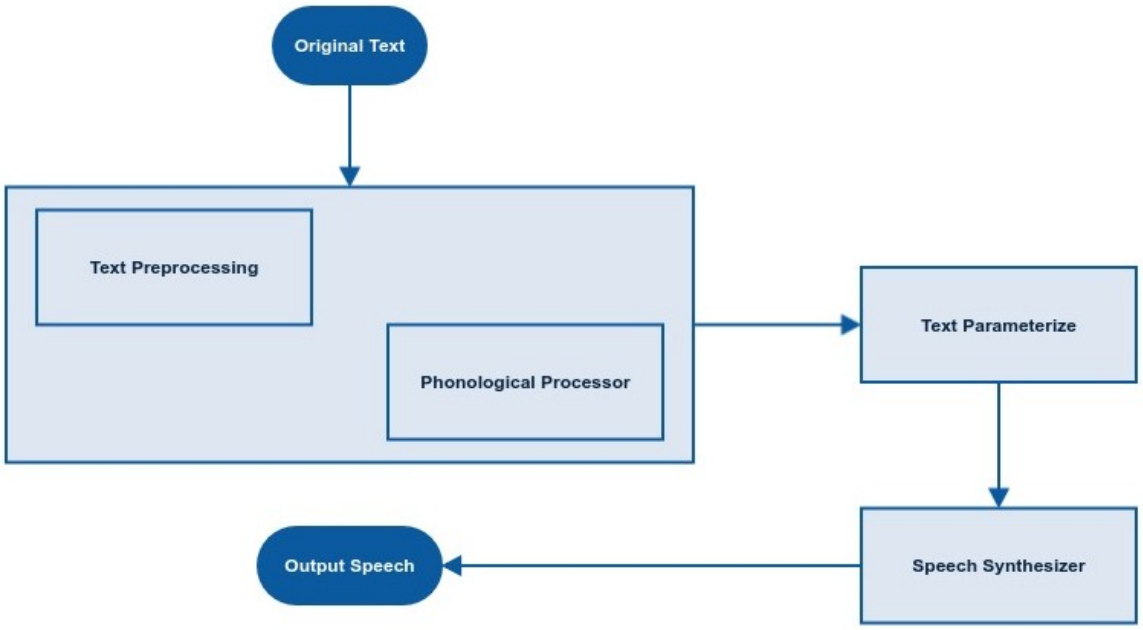

Figure 1-2: Architecture of TTS

The numbers and abbreviations are also processed in this step. In word pronunciation process, words are represented by respective phonetic notations by using word pronunciation dictionary. In phonetic interpretation the duration of phonetic segments, pitch, accents are assigned. Signal generation component of TTS system takes output from all above processes and generate a signal of speech using a function. In [22], text-to-Speech system is divided in two parts. One is called Natural Language Processing unit and other is called Speech Synthesis unit. Natural Language Processing unit preprocess text and converts it into phonetic strings. These phonetic strings are then marked by stress marker and passed to speech synthesis unit which converts it into speech signals.

Chapter 2

Related Work

The TTS conversion is not a new field and people have been working on this field before electronic signal processing techniques. In beginning, people tried to build machines which were used to create human sound. After the development of computers, better systems were built using different techniques. In [12], basic speech synthesizing technique which works by concatenation of small recorded speech segments called phonemes to form complete speech are discussed. Each word is first divided into syllables and then pronunciation for each syllable is concatenated in order to get pronunciation for whole word. This concatenated word has some delay between pronunciations of each syllable which is removed and as a result, final pronunciation of that word is obtained. Problems like Text Preprocessing, Pronunciation and Prosody makes it difficult. In Text Preprocessing, digits and abbreviations are converted to full words. Other problem is guessing correct pronunciation of a word. For example, word lives has different pronunciation in He lives in Lahore and He saved two lives. To create naturalness in sound, stress and intonation are applied to the input text which is also a very complex task.

A rule based technique is designed in [23]. The dataset is developed by extracting 50,000 words from standard Corpus, Corpus of Present-Day Edited American English i.e. Brown corpus [24]. The system gave accuracy of about 93improved system was proposed in [25] and [5].

Dictionary and rule based approach is used in [21] where 1000,000 words were used for training model. A TTS system with prosody and concatenative speech parameters that were extracted through use of probabilistic learning methods in [10]. A formant and Concate-

native synthesis is developed in [26] where small segments of phonemes were concatenated to form whole speech. A training database was used contains about 6,000 phonetically balanced sentences recorded in natural style. In [27], linear regression and unit selection based speech synthesis is designed using ATR Japanese database.

Statistical parametric speech synthesis is another approach which uses parameters to describe speech. In this technique, model is learned from speech data. This technique works better than concatenative technique. In [11], author discussed shortcomings of concatenative synthesis and why Hidden Markov model based technique is better than concatenative.

A Hidden Markov model based Statistical parametric speech synthesis system was developed in [28] where 450 sentences from ATR Japanese Database were used. A similar system is designed in [13] using Hidden Markov Model and evaluated it by taking input from Japanese database and by generating feature vector of those sentences and compared with original speech.

A Hidden Markov model and unit selection based system is proposed in [29] in which 524 sentences from CMU communicator database were taken. The sentences are used to develop system and two algorithms are compared. The result showed unit give high quality results but it uses large data, voice remain fixed whereas HTS gave smooth, stable, various voices but it gives buzzy speech. In [30], Hidden Markov model and rule based approach is applied on voices taken from e-learning courses and online lessons for dataset creation and tested by generating voices and given as input to students to interpret it.

Corpus based approach for Expressive Prosody Modeling is applied in [4] in which manually produced dataset which is not publicly available was used. To evaluate the synthesized speech, the output is given for testing to 32 native English speakers. For bad news, good news and for yes/no the accuracy is 70.2

Tones and Break Indices ToBI are discussed on American English in [31]. Majorly bigram and tri-gram were used to predict the occurrences of particular word or letter. Analysis was performed by multiple techniques. The corpus was divided into following five yes-no questions, either-or questions, other questions (hereafter, wh-questions), exclamations, and other declarative sentences categories for the analysis purpose. Analysis by word frequency count produced not good results and the reason was that the data was of variant types. The results show that professional speakers produces better and informative prosodic events as compared to ordinary speakers.

A TTS system for Azerbhaijani language using concatenative synthesis in [32] where small recordings were concatenated to make speech waveform.

Hidden Markov model based approach is used in [7] to construct speech synthesizer for Xitsonga which is an African language. The system received acceptability of 92.3

TTS system for Fon language is designed in [6] using Multisyn algorithm which consists of Natural Language Processing (NLP) and Digital Signal Processing (DSP) modules. NLP consists of segmentation, Letter-to-Sound conversion and back-off rules module. Back-off rules are applied when input text contains some characters that are not in us know characters. DSP module than choose required unit from database of units are concatenate them to form complete speech signals.

In [33], hybrid text to speech converter is developed by concatenating benefits of HMM based TTS system and waveform based TTS system. System is developed using Matlab, a library of phoneme and their sound is created and waveform of audio and audio itself is generated using these libraries.

Appendix A

Figures

Figure A-1: Armadillo slaying lawyer.

Figure A-2: Armadillo eradicating national debt.

Appendix B

Tables

Table B.1: Armadillos

Armadillos	are
our	friends

Bibliography

[1] Benazir Mumtaz et al. "Break Index (BI) annotated speech corpus for Urdu TTS". In: *Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA), 2016 Conference of The Oriental Chapter of International Committee for*. IEEE. 2016, pp. 22–27.

[2] Miss Prachi Khilari and VP Bhope. "A Review On Speech To Text Conversion Methods". In: *International Journal of Advanced Research in Computer Engineering & Technology* 4.7 (2015).

[3] Dennis H Klatt. "Review of text-to-speech conversion for English". In: *The Journal of the Acoustical Society of America* 82.3 (1987), pp. 737–793.

[4] Ellen Eide et al. "A corpus-based approach to expressive speech synthesis". In: *Fifth ISCA Workshop on Speech Synthesis*. 2004.

[5] Dennis Klatt. "The Klattalk text-to-speech conversion system". In: *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'82.* Vol. 7. IEEE. 1982, pp. 1589–1592.

[6] Theophile K Dagba and Charbel Boco. "A text to speech system for Fon language using multisyn algorithm". In: *Procedia Computer Science* 35 (2014), pp. 447–455.

[7] Ntsako Baloyi. "A text-to-speech synthesis system for Xitsonga using hidden Markov models". PhD thesis. University of Limpopo (Turfloop Campus), 2012.

[8] *Top 30 Languages by Number of Native Speakers*. URL: http://www.vistawide.com/languages/top_30_languages.htm.

[9] ABDUL MANNAN Saleem et al. "Urdu consonantal and vocalic sounds". In: *CRULP Annual Student Report* (2002).

[10] Xuedong Huang et al. "Whistler: A trainable text-to-speech system". In: *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference On*. Vol. 4. IEEE. 1996, pp. 2387–2390.

[11] Thomas Merritt and Simon King. "Investigating the shortcomings of HMM synthesis". In: *Eighth ISCA Workshop on Speech Synthesis*. 2013.

[12] N Swetha and K Anuradha. "Text-to-speech conversion". In: *Int J Adv Trends Comput Sci Eng* 2.6 (2013), pp. 269–278.

[13] Keiichi Tokuda et al. "Speech parameter generation algorithms for HMM-based speech synthesis". In: *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*. Vol. 3. IEEE. 2000, pp. 1315–1318.

[14] Nobuyoshi Kaiki. "Linguistic properties in the control of segmental duration for speech synthesis". In: *Talking Machines: Theories, Models, and Designs* (1992), pp. 255–263.

[15] Michael D Riley. "Tree-based modeling of segmental durations". In: *Talking machines* (1992), pp. 265–273.

[16] Alan W Black, Heiga Zen, and Keiichi Tokuda. "Statistical parametric speech synthesis". In: *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*. Vol. 4. IEEE. 2007, pp. IV–1229.

[17] Heiga Zen et al. "The HMM-based speech synthesis system (HTS) version 2.0." In: *SSW*. Citeseer. 2007, pp. 294–299.

[18] Anand Arokia Raj et al. "Text processing for text-to-speech systems in Indian languages." In: *SSW*. 2007, pp. 188–193.

[19] Hasan Kabir et al. "Natural language processing for Urdu TTS system". In: *Multi Topic Conference, 2002. Abstracts. INMIC 2002. International*. IEEE. 2002, pp. 58–58.

[20] Sarmad Hussain. "Phonological Processing for Urdu Text to Speech System". In: *Yadava, Y, Bhattarai, G, Lohani, RR, Prasain, B and Parajuli, K (eds.) Contemporary issues in Nepalese linguistics* (2005).

[21] Mark Y Liberman and Kenneth W Church. "Text analysis and word pronunciation in text-to-speech synthesis". In: *Advances in speech signal processing* (1992), pp. 791–831.

[22] H. R. Basit and S Hussain. *Text Processing for Urdu TTS System*. Poster presentation in Conference on Language and Technology 2014 (CLT 14), Karachi, Pakistan. 2014.

[23] Honey S Elovitz et al. *Automatic translation of English text to phonetics by means of letter-to-sound rules*. Tech. rep. NAVAL RESEARCH LAB WASHINGTON DC, 1976.

[24] H Ku, WN Francis, et al. "Computational analysis of present-day {A} merican {E} nglish". In: (1967).

[25] Rolf Carlson, B Granstrom, and Sheri Hunnicutt. "A multi-language text-to-speech module". In: *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'82.* Vol. 7. IEEE. 1982, pp. 1604–1607.

[26] Xuedong Huang et al. "Recent improvements on Microsoft's trainable text-to-speech system-Whistler". In: *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*. Vol. 2. IEEE. 1997, pp. 959–962.

[27] Andrew J Hunt and Alan W Black. "Unit selection in a concatenative speech synthesis system using a large speech database". In: *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*. Vol. 1. IEEE. 1996, pp. 373–376.

[28] Takayoshi Yoshimura et al. "Duration modeling for HMM-based speech synthesis". In: *Fifth International Conference on Spoken Language Processing*. 1998.

[29] Keiichi Tokuda, Heiga Zen, and Alan W Black. “An HMM-based speech synthesis system applied to English”. In: *IEEE Speech Synthesis Workshop*. 2002, pp. 227–230.

[30] Hideto D Harashima. “Review of ”VoiceText””. In: *Electronic Journal of Foreign Language Teaching* 3.1 (2006), pp. 131–135.

[31] John F Pitrelli. “ToBI prosodic analysis of a professional speaker of American English”. In: *Speech Prosody 2004, International Conference*. 2004.

[32] fffdfffdR Aida–Zade, C Ardil, and AM Sharifova. “The main principles of text-to-speech synthesis system”. In: *International Journal of Signal Processing* 6.1 (2010), pp. 13–19.

[33] Mohd Bilal Ganai and Er Jyoti Arora. “Text-to-Speech Conversion”. In: (2016).

[34] Rolf Carlson and B Granstrom. “A text-to-speech system based entirely on rules”. In: *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP’76*. Vol. 1. IEEE. 1976, pp. 686–688.