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

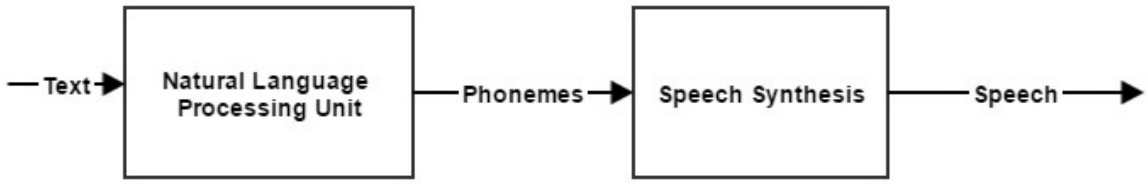
Chapter 1

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.



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.

Chapter 2

Related Work

Let's cite! The Einstein's journal paper [**westwood1998validation**] and the Dirac's book [**dirac**] are physics related items. [**ryan2001narrative**] virtual

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

References

[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. "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.

[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.