A DNN-BASED TEXT-TO-SPEECH SYSTEM FOR HAUSA: AN UNDER-RESOURCED LANGUAGE

ABUBAKAR AHMAD ALIERO

FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2017

A DNN-BASED TEXT-TO-SPEECH FOR HAUSA: AN UNDER-RESOURCED LANGUAGE

ABUBAKAR AHMAD ALIERO

DISSERTATION SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SOFTWARE ENGINEERING

FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2017

UNIVERSITY OF MALAYA
ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: Abubakar Ahmad Aliero

Matric No: WGC 150025

Name of Degree: Master of Software Engineering

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

A DNN-based Text-to-Speech System for Hausa: An Under-Resourced

Language Field of Study: Human Computer Interaction

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date:

Subscribed and solemnly declared before,

Witness's Signature Date:

Name:

Designation:

ABSTRACT

In recent years, speech technology has gained a tremendous improvement in term of its application and development. Speech technology such as machine translator, automatic speech recognition system and speech synthesis system are the state-of-the-art in today's technology. TTS system or artificial speech development during the last few decades aims at gradual improvement in the intelligibility and naturalness. A Text-to-Speech system is a system that generates speech output from a given input text. TTS system has many different applications for many different users, but more specifically are the visually impaired and the illiterates. Some of the major application areas of speech synthesis system are document reader, speech translator, mobile read-aloud applications (such as google map reader) and announcement system. Speech synthesis system serves as an assistive tool for disabled, which is used for reading online text/information and as an automatic learning system for children.

Despite the potential benefits of TTS system, it is language dependent and has yet to be developed for many of the languages around the world, which is mostly due to the lack in the necessary resources. Languages that is lacking in the necessary resources are referred as under-resourced language. Hausa is one of the under-resourced languages that lacks in the resources for developing a TTS system. The aim of this research is to develop a state-of-the-art TTS system for Hausa, an under-resourced language, using minimal resources. Several techniques have been introduced by researchers for developing TTS system for under-resourced languages, such as speaker adaptation, cross-lingual adaptation, bootstrapping, and etc. Currently, the state-of-the-art TTS technology is the Deep Neural Network (DNN)-based speech synthesis system which is only available for selected well-resourced languages like English, Arabic etc. The DNN-based speech synthesis system is the most advanced system that offers the highest intelligibility and naturalness as compared to the existing systems. Using the English resources as the basis, a DNN-based speech synthesis system is developed for Hausa with minimal resources by adopting the cross-lingual technique. The developed system was tested for intelligibility and naturalness using native Hausa speakers. The result of the developed system is 4.20 out of 5 in terms of naturalness and 4.10 out of 5 in terms in intelligibility, which is better than the existing techniques used for the development of TTS systems for under-resourced languages.

ABSTRAK

Dalam tahun-tahun kebelakangan ini, teknologi pertuturan mendapat peningkatan yang besar dari segi aplikasi dan pembangunan. Teknologi ucapan seperti penterjemah mesin, pengecaman pertuturan automatik dan sistem sintesis pertuturan adalah teknologi sangat terkini. Pembangunan sistem Text-to-Speech (TTS) atau pertuturan tiruan bagi beberapa dekad yang lalu adalah bertujuan untuk peningkatan beransur-ansur dalam kejelasan dan keaslian. Sistem TTS adalah sistem yang menjana output pertuturan daripada teks yang diberikan. Sistem TTS mempunyai beberapa aplikasi yang berbeza untuk ramai pengguna yang berbeza, tetapi lebih khusus adalah untuk orag cacat penglihatan dan buta huruf. Beberapa aplikasi utama sistem sintesis pertuturan adalah pembaca dokumen, penterjemahan ucapan, aplikasi baca dengan jelas mudah alih (seperti peta pembaca google) dan sistem pengumuman, sistem sintesis pertuturan berfungsi sebagai alat bantuan untuk orang kurang upaya, yang digunakan untuk membaca teks/maklumat atas-talian, dan sistem pembelajaran automatik untuk kanak-kanak.

Walaupun terdapat banyak faedah bagi sistem TTS, ia sangat bergantung kepada bahasa dan tidak dibangunkan untuk kebanyakan bahasa di seluruh dunia, yang disebabkan oleh kekurangan sumber yang diperlukan untuk pembangunan sistem bagi bahasa berkenaan. Hausa adalah salah satu bahasa yang kekurangan sumber-sumber untuk membangunkan sistem TTS. Tujuan kajian ini adalah untuk membangunkan satu sistem TTS berasakan teknik terkini untuk bahasa Hausa, dengan menggunakan sumber pada kadar yang minimum. Beberapa pendekatan telah diperkenalkan oleh penyelidik untuk membangunkan sistem TTS untuk bahasa kekurangan sumber seperti ‘speaker adaptation’, ‘cross-lingual adaptation’, ‘bootstrapping’ dan sebagainya. Pada masa ini, teknik TTS terkini adalah berasakan Deep Neural Network (DNN) yang hanya telah digunakan untuk pembangunan TTS bagi bahasa yang mempunyai banyak sumber seperti Inggeris, Bahasa Arab dan lain-lain. Sistem TTS yang berasaskan DNN adalah sistem yang paling maju yang menawarkan kejelasan dan keaslian yang paling tinggi berbanding dengan sistem yang sedia ada. Menggunakan sumber-sumber bahasa Inggeris sebagai asas, sistem TTS berasaskan DNN dibangunkan untuk Hausa dengan sumber yang minimum dengan mengguna pakai teknik 'cross-lingual'. Sistem yang dibangunkan telah diuji untuk kejelasan dan keaslian ucapan menggunakan responden-Hausa asli. Hasilnya adalah 4.20 daripada 5 dari segi keaslian dan 4.10 daripada 5 dari segi kejelasan ucapan sintesis, iaitu keputusan yang lebih baik daripada pendekatan yang digunakan untuk sistem TTS sedia ada.

ACKNOWLEDGEMENTS

I would like to express my sincere and endless gratitude to Allah S.W.T for given me health, strength, patience and ability to complete this dissertation. I sincerely thanks my supervisor, Dr. Mumtaz Begum Binti Peer Mustafa for her extensive support throughout the period of this research work, without her continued, enthusiastic and comprehensive guidance; it would not have been possible for me to write this page today. Starting from her energetic discussions to stimulating ideas about research, she was successful to induce a high level of liveliness to her students. Working closely with her for almost the period of this research, I learnt a lot about hardworking and time management.

I would also like to thank Dr. Tijjani S. Almajiri and Dr. Zubairu Malah who are always there to assist and guide me throughout the period of resource gathering. My sincere appreciation goes to Shehu Habibu, Baba Adamu and Dalhatu Muhammed Whose their voice were used for the conduct of this research. A very special thanks goes to Dr. Muhammad Garba and Aminu Zimit who assisted in monitoring the recording process in both London and Kano respectively. I also thanks all my colleague at the faculty and university in general.

This acknowledgement is incomplete without thanking my brother Mukhtar and other family members for their support and prayers. The patience, support, prayers of my wife for the whole period of this study is well acknowledged.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

ANN	:	Artificial Neural Network
ASR	:	Automatic Speech Recognition
BBC	:	British Broadcasting Corporation
CART	:	Classification And Regression Tree
CV	:	Consonant Vowel
CVC	:	Consonant Vowel Consonant
CVV	:	Consonant Vowel Vowel
DNN	:	Deep Neural Network
DRT	:	Diagonostic Rhyme Test
DSP	:	Digital Signal Processing
F0	:	Fundamental Frequency
GB	:	Gigabyte
GHz	:	Gigahertz
GMM	:	Gaussian Mixture Model
GPS	:	Global Positioning System
G2P	:	Grapheme-to-Phoneme
HDD	:	Hard Disk Drive
HMM	:	Hidden Markov Model
HTK	:	HMM-toolkit
IPA	:	International Phonetic Alphabet
IIUM	:	International Islamic University Malaysia
LSTM	:	Long Short Term Memory
LTS	:	Letter-to-Sound
MAP	:	Maximum *a posteriori*

MDN	:	Mixture Density Network
MFCC	:	Mel-frequency Cepstral Coefficient
MLLR	:	Maximum Like-lihood Linear Regression
MLSA	:	Mel Log Spectrum Approximation
MOS	:	Mean Opinion Score
MRT	:	Modified Rhyme Test
MSD	:	Multi-Space Probability Distribution
NLP	:	Natural Language processing
POS	:	Part of Speech
RFI	:	Radio Franc Internationale
RLAT	:	Rapid Language Adaptation Toolkit
RNN	:	Recurrent Neural Network
SPSS	:	Statistical Parametric Speech Synthesis
TIMIT	:	Texas Institute and Massachusetts Institute of Technology
TTS	:	Text-to-Speech
UM	:	University of Malaya
UPM	:	University Putra Malaysia
U-RL	:	Under-Resourced Language
US	:	United State
USS	:	Unit Selection Synthesis
VOA	:	Voice of America
WER	:	Word Error Rate

LIST OF APPENDICES

CHAPTER 1: INTRODUCTION

Speech technology has achieved a remarkable progress in recent years with the development of many speech-based applications in the field of communication, transport, industries, weather, and so on. There are three major components of speech technology, which are the Automatic Speech Recognition (speech-to-text), Speech Synthesis System (text-to-speech) and Machine Translation (text-to-text and speech-to-speech) system. Speech technology has increased the number of computer’s users to include individuals such as the visually impaired and illiterates.

Verbal-based communication is the most efficient and perfect way of communication. As such applying speech-based interaction between man and machine can increase the effectiveness of the interaction that is currently limited to the visual and mechanical form of interaction. One of the speech-based technologies is the TTS system, which generates human-like speech from written text input. There are many areas of applications for TTS system including education, communication as well as an assistive tool for individuals with physical impairment.

TTS system plays a very important role in children learning, as well as for learning the second language. TTS system can be used as an automatic learning system for children that can help them in learning words and the correct pronunciation. TTS system can also be very useful for many applications such as GPS reader, public transit system, weather report, and so on.

TTS system has enabled the participation of individuals such as the blind and the illiterates in using computer technology, which greatly improves their lives. Before the development of TTS system, the blind and illiterates cannot access the information from the computers and the Internet. The current visual and mechanical interactions are disadvantageous to the blind and the illiterates, due to the inability to read the information presented on the computer screen. Visual-based information such as e-books, online magazines, and SMS on mobile phones are some of the information that they cannot access. TTS system can also enable the current users (regular sighted users) of computers and smartphones to conveniently listen to the synthesized speech from various sources of written information when they are busy doing other activities such as driving. TTS system allows them to access to information with less effort and without straining their eyes when reading text on a small screen.

1.1 Research Background

TTS system is increasingly accepted as the assistive tool for people suffering from visual impairment and the illiterate, where their physical and socioeconomic limitations make them unable to read the written text. TTS system has greatly enhanced the lives of those individuals (Isewon, Oyelade, & Oladipupo, 2014).

The process of converting the text input into speech output comprises of two parts, which are the high-level synthesis and the low-level synthesis. High-level synthesis is the transformation of the text input into phonetic or other forms of linguistic representation, while the low level synthesis is the transformation of phonetic and prosodic information into speech waveforms (Isewon et al., 2014).

The two major units of TTS system are the Natural Language Processing (NPL) and the Digital Signal Processing (DSP).

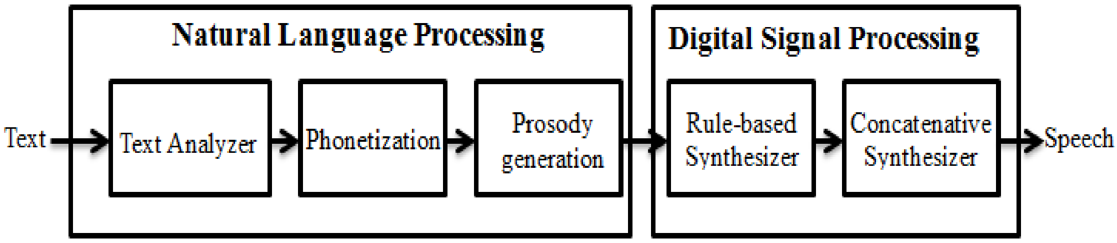

Natural Language Processing
Digital Signal Processing
Text
Text Analyzer
Phonetization
Prosody generation
Rule-based Synthesizer
Concatenative Synthesizer
Speech

Figure 1.1: Functional Diagram of Text-To-Speech System (Rashad, El-Bakry, Isma'il, & Mastorakis, 2010).

The phonetic and linguistic information generated by the Natural Language Processing (NLP) unit is used by the Digital Signal Processing (DSP) unit to generate the waveform with appropriate stress, rhythm, and intonation. The DSP generates the speech waveform by concatenating the pre-recorded speech units or by applying a speech acoustic model. The more recent state-of-the-art TTS systems based on the Hidden Markov Models (HMMs) and Deep Neural Networks (DNNs) have increased the effectiveness of the DSP unit to synthesize synthetic speech with acceptable quality when the development has adequate size and high-quality speech resources. There are two major attributes of TTS system, which are the naturalness and intelligibility.

- **Naturalness**

Naturalness is the measures of the degree of similarities between the human speech and the synthesized speech. Synthetic speech is highly natural when the listeners cannot distinguish between the synthetic and the human speech.

- **Intelligibility**

Intelligibility is the measures of the ability of human listeners to correctly comprehend the synthesized speech. The TTS system is intelligible if the listeners can correctly understand the synthesized speech with the intended meaning.

1.1.1 Natural Language Processing (NLP)

The NLP unit converts text input into a symbolic representation, where it is responsible for converting the written text into its corresponding phonetic transcription together with the desired intonation and rhythm (Dutoit, 1997). NLP unit consists of three phases, which are text analyzer, phonetization, and prosody generation. One of the examples of the existing natural language processing engine is the Festival speech synthesizer.

The NLP unit is language dependent and processes the textual-based information of a particular language, including the orthography, phonology or morphology. As such, the intelligibility and naturalness of a TTS system for a particular language depend on the performance of the NLP units, particularly on its ability to process the text input to its equivalent phonetics representation.

- **Text Analyzer**

Text analyzer is responsible for analyzing the text input text that involves several stages. For the first stage, numbers, acronyms, and abbreviations are converted into full text, and then decompose the input sentences into groups of words. The first stage is known as the pre-processing stage. The second stage is the morphological analysis, where words of the sentence that have been analyzed are categorized into possible parts of speech, while compound words are divided into their basic unit before being analyzed.

The third stage of a text analyzer is the contextual analysis module. In this stage, words are considered into their context, which allows the reduction to the list of the possible part-of-speech categories to restrict the number of highly probable occurrence, given to the corresponding possible parts of speech of neighboring words. The fourth stage is the syntactic-prosodic parser, which determines the text structure that tends to be closer to the prosodic realization of the input sentence (Onaolapo, Idachaba, Badejo, Odu, & Adu, 2014).

- **Phonetization**

The Letter-To-Sound (LTS) module is responsible for the automatic determination of the phonetic transcription of the text input. This process is more than just generating the pronunciation of the words from the dictionary as many of the words from the input text can have different phonetic transcription, which depends on the context such as the location and meaning.

- **Prosody Generation**

Prosody generation process focuses on the precise section of a sentence, such as an emphasis on a specific syllable, and so on. This process also helps to segment sentences into smaller units comprising of groups of words and syllables and also to identify the relationship between those units. The prosody generator is responsible for generating the various aspect of speech including tone, accent, and emphasis of a sentence (Onaolapo et al., 2014).

1.1.2 Digital Signal Processing (DSP)

The DSP unit is responsible for the conversion of the symbolic representation generated by the NLP unit into the audio signal or synthesized speech. The DSP unit can be categorized as into several techniques, such as rule-based, concatenative or statistical parametric synthesis. The DSP is important in attaining the naturalness and the intelligibility of a TTS system by generating the accurate acoustic model and provide complex dependencies between linguistic and acoustic features (Zen, 2015).

- **Rule-based Synthesis**

In rule-based speech synthesizer, speech is generated from the dynamic modification of several speech parameters digitally. These parameters are modified digitally to create an artificial speech waveform. An example of those parameters is fundamental frequency, voicing, and noises. The development of the rule-based speech synthesizer is time-consuming as it requires complex rules to be created and any errors in those rules will result in poor quality synthesis (Onaolapo et al., 2014).

- **Concatenative Synthesis**

Concatenative synthesizers joined together pieces of human recorded speech extracted from a speech database. Concatenative speech synthesizer uses different speech units such as diphone, syllable, and even a complete word. The database size depends on the type of speech unit used. In order to increase the naturalness of the synthesizer, a large database of source speech is required. The two main types of concatenative speech synthesizers are the Unit selection synthesis and Diphone-based synthesizer.

Unit selection synthesizer uses large databases of recorded speech. The database contains recorded utterance that has been segmented into many different speech units such as phones, diphones, half-phones, syllables, morphemes, words, phrases, and sentences. Typically, the division of speech into segments is performed using specially modified speech recognizer that will “forced align" the recorded speech, which is then followed by some manual corrections, using visual representations tools such as the waveform and spectrogram (Black, 2002). The unit selection concatenative synthesizer finds the sequence of speech units that matches the best sound or phrase to be synthesized, where the selection is performed according to the descriptors of the units, which are the characteristics extracted from the source sounds, or higher level descriptors attributed to them (Schwarz, 2006).

Diphone-based synthesizer uses speech database that only contains the diphone units (sound-to-sound transitions) of a particular language. The number of diphones depends on the phonotactics of the language. For diphone-based synthesizer, only one sample of each diphone is found in the speech database. At runtime, the target prosody of a sentence is superimposed on these minimal units by means of techniques such as linear predictive coding, PSOLA or MBROLA. The quality of the generated speech is usually poorer than the unit-selection synthesizer but more natural-sounding than the output of rule-based synthesizer such as the formant synthesizers (Isewon, Oyelade, & Oladipupo, 2014).

- **Statistical Parametric Synthesis**

Statistical Parametric Speech Synthesis (SPSS) is the state-of-the-art synthesizer and is dominating the speech technology research and development at this present moment. SPSS uses a set of parametric representation of speech to generate a speech waveform. This parameters (spectral and excitation) are extracted from the recorded human speech database to generate speech the acoustic model. SPSS can generate highly natural and flexible speech from relatively small trained speech acoustic model. Presently, there are two types of statistical parametric speech synthesizers, which are based on the Hidden Markov Model (Zen, Tokuda, & Black, 2009), and Deep Neural Network (Zen, Senior, & Schuster, 2013).

1.1.3 Text-To-Speech for Under-Resourced Languages

Krauwer (2003) and Berment (2004) introduced the term “under-resourced languages” to reflect that are lacking in some or all of the resources need for the development of a TTS system, such as the difficulty in obtaining representative data for acoustic and language model, lack of special phonological systems, lack of linguistics expertise, word segmentation problems, lack of electronic resources for speech processing, unwritten language, and limited presence of the web sources (Besacier, Barnard, Karpov, & Schultz, 2014). Examples of under-resourced languages are Tamil, Assamese, Kannada, Yoruba, Afrikaans, Marathi, Bangladeshi Bangla, Bahasa Melayu, Taiwanese, Swahili, and so on.

Hausa is an under-resourced language that belongs to the Chadic family of languages of the Afro-Asiatic phylum, which spoken by more than 50 million people in West Africa as their mother tongue, second language, and lingua franca. The Hausa language is spoken mainly in the Sahel region of Africa, which consists of Northern Nigeria, Southern Niger, Southern Chad, Northern Cameron, and the Central Republic of African. Hausa is also spoken in some of the Western countries like Germany. The Hausa language is spoken by almost 29 million indigenous (Northern Nigeria) and 18 million non-indigenous speakers (Niger, Ghana, Cameroon, and Benin) The Hausa language has been in existence since before the period of colonization. At that time, it was written in Arabic script, which is called the Ajami or Hausa Arabic script (Philips, 2004). Hausa consists of two major dialects, which are, the Eastern Hausa (e.g. Kano Hausa's, Zinder Hausa's, Hadeja Hausa's, and e.t.c), and the Western Hausa (e.g Sokoto Hausa's, yauri Hausa's, zamfara Hausa's, and e.t.c). The Eastern Hausa dialect is considered to be the standard Hausa, which is used as a system in writing The Hausa language. Despite the variation of language between the Western and Eastern dialects of Hausa, they enjoy a high degree of mutual intelligibility. The major variation between the dialects is the difference in pronunciation, vocabulary, and to a lesser extent morphology.

The Hausa language uses Boko (western education) as the official method of writing system which was introduced by the British during colonization. The second Hausa script which is known as Ajami uses Arabic letters in the system of writing and is mainly used by Arabic-Speaking Hausa scholars and some traditional rulers that lack the western education. The Hausa language has 22 consonants, 5 vowels and 3 diphthongs (Schlippe, Djomgang, Vu, Ochs, & Schultz, 2012). Table 1.1 shows the lists of Vowels, Diphthongs, and Consonant for Hausa. The details of Hausa Language is further described in chapter 4, section 4.1.

Table 1.1: Hausa Vowels, Diphthongs and Consonant

Vowels	a, i, o, u, e
Diphthongs	ai, au ui
Consonants	b c d f g h j k l m n r s t w y z kw gw fy ts sh gy

1.2 Research Motivation

In recent years, researchers have aimed to develop a TTS system for many of the under-resourced languages in order to reduce the technology gap between the regular users and the visual impaired/illiterates. TTS system is language dependent as TTS system of a language cannot be applied to another language without any modification. TTS system leaves the positive impact on the life of many people such as the visually impaired, and is used in many applications such as transport schedule reading system, google map reader and so on.

Despite the popularities of Hausa, it lacks with the proper representation in electronic resources for speech processing, limited presence on the web, and it lacks researchers in the field of speech technology. The lack of TTS system for Hausa has motivated the researcher to develop a TTS system for Hausa using state-of-the-art technique with adequate intelligibility and naturalness.

1.3 Research Problems

Despite the fact that speech can be a successful medium of man-machine interaction, speech technology related applications are predominantly language dependent, which implies that the systems or application of one language cannot be suitable for another language without any modification or adaptation. The lack of the progress in TTS system for the under-resourced languages is mainly attributed to the non-availability of resources such as recorded speech database, speech technology expertise, and funding.

The lack of speech technology experts such as researcher, phonologist, and machine learning experts for Hausa has resulted in the lack of the development for TTS system for Hausa. On top of that, resources such as speech data, transcription, pronunciation dictionary, labels, and letter-to-sound rules are crucial for the development of a TTS system for a new languages. All these are some of the issues that prevent the development of TTS system for Hausa. Development of a TTS system for Hausa requires the identification of suitable techniques that have the ability to synthesize Hausa speech with acceptable intelligibility and naturalness with the use of minimal resources.

The speech database (also known as speech corpus) consists of speech audio and text transcriptions. Speech database is one of the critical components in the development of a TTS system. The non-availability of speech database hinders the development of the TTS system for under-resourced languages. In TTS system development, it is necessary to ensure that all possible phonemes and phoneme combinations of a particular language are included. The text prepared for the recording should adequately cover the phoneme representation of a particular language (Navas, Hernaez, & Luengo, 2006). A good quality speech database ensures the quality of the speech acoustic model in order to synthesize speech with the high degree of intelligibility and naturalness.

Speech technology expert is one of the most important resources that contribute to the development of TTS system for a new language. Many under-resourced languages lack in the experts for the development of TTS system for their own language. There are several reasons for this, ranging from the lack of interest among the native speakers or the speakers are not having the adequate knowledge and skills due to inability to have access to the skills and knowledge required for developing a TTS system (Molapo et al., 2014).

1.4 Research Objectives

1. To accumulate resources for the development of Text-To-Speech system for Hausa.
2. To identify suitable state-of-the-art technique(s) for the development of TTS system for Hausa with adequate intelligibility and naturalness.
3. To develop and evaluate the Text-To-Speech system for Hausa using the identified techniques and evaluation methods.

1.5 Research Questions

The following research questions are established and serve as guidance to conduct this research at various stages:

- RQ1. What are the issues that limit the development of Text-to-Speech System for Hausa?
- RQ2. What are the minimal resources needed for developing the Hausa Text-To-Speech System?
- RQ3. What are the suitable state-of-the-art technique(s) for the development of Hausa Text-To-Speech system?

1.6 Research Scope

The focus and scope of this research are to develop a TTS system for Hausa Language using standard (Eastern) Hausa words and sentences only. To reduce the speech variation, only the speeches from a male speaker is considered in this research as male speech has low variations as compared to female speech, and any of the resources not available for Hausa will be borrowed from the English language resources due to it's similarities with Hausa language especially in the segmental phonemes.

1.7 Research Methodology

This section provides a brief introduction to the research methodology adopted in this research. This research consists of four main stages, which are, the literature review, data accumulation, development and evaluation.

- **Literature Review**

The first stage of this research is to conduct the review of the existing literature to identify the issues and the possible solution towards the development of a TTS system for the under-resourced language, development of TTS system for the under-resourced language, and the evaluation method to be used for evaluating the developed TTS system. The purpose of the review is to identify a suitable state-of-the-art technique for the development of the TTS system for the under-resourced language. This review is also to identify techniques for resource accumulation, and evaluation method for the developed system.

- **Data Accumulation**

This stage involves the data accumulation such as text collection, text corpus development, recording of selected sentences, transcription, and the development of pronunciation dictionary.

- **Development**

The development stage involves the development of the Hausa acoustic model using the identified technique and accumulated resources. The development consists of features extraction, generation of the time-align phone transcription, normalization, and forced-alignment. The performance of the training data is also determined at this stage.

- **Evaluation**

The last stage of this research is the evaluation of the developed system and analysis of the results. This process involves the selection of sentences for the evaluation, target respondents for the evaluation of the developed system in term of intelligibility and naturalness. The findings from the analysis are then compared with the result of the existing TTS system developed for other under-resourced languages.

1.8 Organization of the Research

The rest of this dissertation is organized as follows;

Chapter 2 discusses the techniques used for the development of Text-to-Speech system as well as their merit and demerit. The techniques/method used for the development of TTS system for under-resourced language is also explained in this chapter. This research also reviews the methods used for resource accumulation and systems development for many of the under-resourced languages. This chapter further discusses the various TTS system developed for under-resourced languages, their performance, and method of evaluation.

Chapter 3 discusses and explained the research methodology adopted in the conduct of this research. It also highlights the justification of the selected methods and technique employed in this research for the resource accumulation and the development of TTS system for The Hausa language.

Chapter 4 describes the structure of the Hausa language, which includes resource accumulation such as text corpus, and the development of speech corpus, which includes speech recording, transcription, and segmentation and labeling.

Chapter 5 discusses the development processes of Hausa TTS system using the proposed techniques and accumulated resources.

Chapter 6 discusses the evaluation process of the developed Hausa TTS system. The result of the evaluation and findings are also discussed in this chapter.

Chapter 7 Summarizes and concludes the findings of this research. It also discusses some of the possible future works for further improvements to the TTS system for the under-resourced languages.

CHAPTER 2: LITERATURE REVIEW

This chapter presents the findings from the review of the existing literature that discusses the issues, and the possible techniques that are suitable for the development of a TTS system particularly for under-resourced language, as well as the resources needed for the development of TTS system for the under-resourced language. This chapter also presents the evaluation methods used to evaluate a TTS system.

2.1 Human Speech Production Mechanism

“To understand the speech production, it is important to know and understand the mechanism of the human speech production and several characteristics of the speech signal. Human speech is produced by a series of vocal organ, with lungs and the diaphragm as the main source of energy for the production of sound. The airflow is forced through the glottis between the vocal cords and the larynx to the three main cavities of the vocal tract, which are, pharynx, oral, and nasal cavities. From the oral and nasal cavities, the airflow exists through the nose and mouth respectively. The V-shape opening at the vocal cords called the glottis and is the most important sound source in the vocal system. The vocal cords may act in several different ways during the speech.

The most important function of the vocal cord is to modulate the airflow by rapid opening and closing, causing buzzing sound from which vowels and voiced consonants are produced. The fundamental frequency of vibration depends on the mass and tension, and it is about 110Hz, 200Hz and 300Hz for men, women, and children respectively. For the stop consonants (b, d, t etc.), the vocal cords may act suddenly from a completely closed position, in which it cut the airflow completely and to a totally open position, producing a light cough or a glottal stop.

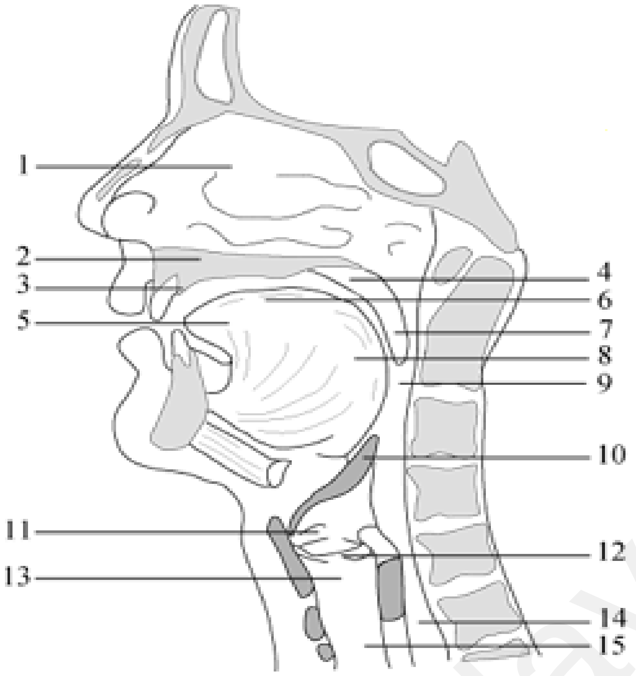


Figure 2.1: Human Speech Production System.

1. Nasal Cavity, 2. Hard palate, 3. Alveoral ridge, 4. Soft palate, 5. Tip of the tongue, 6. Dorsum, 7. Uvula, 8. Radix, 9. Pharynx, 10. Epiglottis, 11. False vocal cords, 12. Vocal cords, 13. Larynx, 14. Esophagus 15. Trachea (Karjalainen, 1999).

The unvoiced consonants may be completely open and an intermediate position may also occur with some phoneme. The pharynx connects the larynx to the oral cavity and it has almost fixed dimensions, but its length may change slightly by raising or lowering the larynx at one end and the soft palate at the other end. The soft palate also isolates or connects the route from the nasal cavity to the pharynx. At the bottom of the epiglottis and false vocal cords, the food is prevented from reaching the larynx, and to isolate the esophagus acoustically from the vocal tract. The epiglottis, the false vocal cords, and the vocal cords are closed during swallowing and open during normal breathing" (Keller, 1995). This complex process of human speech production has made it not easy to be imitated by the TTS system.

2.2 The Usefulness and Importance of TTS System

TTS system is an artificial production of human speech. In the past, specific audio books were used, where the content of the book is read into the audio tape by a reader. It is clear that making such spoken copy of any large book takes several months and is very expensive. TTS system can be important and useful to the many different layers of the society, the most important uses of a TTS system is to serve as a communication and reading aid for the visually impaired people.

Many TTS system applications have been developed and deployed in embedded devices to assist many different users not limited to the visually impaired. An example of such devices includes cell phone, toys, GPS systems, and large scale systems for directory assistance and customer care. TTS system has generally increased the participation of the visually impaired in the field of computer technology.

2.3 Types of Speech Synthesis Technique

TTS system is a process of generating speech from text input by a computer or smartphone. TTS system generates human-like speech not from pre-recorded speech (e.g. queue management system), but either from techniques such as rule-based or concatenative techniques. Many of the existing TTS systems have been developed using several different techniques such as Rule-based synthesis (Articulatory synthesis and Formant synthesis), Concatenative Synthesis (Unit Selection Synthesis and Diphone synthesis), and HMM Speech Synthesis (Rashad et al., 2010).

2.3.1 Rule-based Synthesis

The rule-based synthesis generates artificial speech through the dynamic modification of several speech parameters, such as fundamental frequency, voicing and so on. The two major rule-based synthesis are articulatory and formant synthesis.

- **Articulatory Synthesis**

Articulatory Synthesis is one the rule-based synthesis that uses a set of rules in production of speech. Articulatory synthesis generates speech by direct simulation of the human speech or voice in which it models the articulatory behavior of human. Articulatory synthesis generates the most similar and high quality synthesized speech, similar to the human voice. The speech generated in articulatory synthesis is more naturalness but it is the most difficult method to implement. Some of the mechanisms that regulate the articulation include lip aperture, lip protrusion, tongue tip position, tongue tip height, tongue position, and tongue height.

Although articulatory synthesis provides a very similar human speech, it has some difficulty, which is the difficulty in obtaining data for articulatory model (this data is usually derived from x-ray photography; X-ray data do not characterize the masses or degrees of freedom of the articulators), and the difficulty in finding a balance between a highly accurate model and a model that is easy to design and control. In general, the results of articulatory synthesis (e.g. CASY) (Iskarous, Goldstein, Whalen, Tiede, & Rubin, 2003) are not good as the formant synthesis or the concatenative synthesis (Rashad et al., 2010).

- **Formant Synthesis**

Unlike articulatory synthesis, Formant synthesis synthesizes speech using some instruments that are governed by a set of rules. A good example of Formant TTS system is Klatt formant synthesizer (Anumanchipalli et al., 2010).

Formant synthesis is based on the source-filter model of speech production. The generated speech waveform does not use any natural recorded speech, as it is derived from some parameters (fundamental frequency, amplitude of voicing, open quotient, nasal pole frequency etc.) (Figueiredo, Imbiriba, Bruckert, & Klautau, 2006), the sound source for vowels is a periodic signal with a fundamental frequency and for unvoiced consonants, a random noise generator is used. For formant synthesis, estimated frequency is used to synthesize speech and this frequency makes the sound distinct models the pole frequencies of speech signal.

Formant synthesizer is categorized into two, which are cascade and parallel synthesizer. For cascade formant synthesizer, a series connection of resonators exists where each resonator output is fed into the next one. On the other hand, for the parallel formant synthesizer, each resonant is modeled separately and the source signal is fed through each resonant which is then all summed together. The parallel configuration has an amplitude controlling each formant. The series connection of resonators in cascade arrangement is simpler compare to that of parallel arrangement. The cascaded structure produces a better sound for non-nasal words while the parallel structure is good in producing nasals and fricatives utterance (Rashad et al., 2010). The formant synthesizer produces a monotonic and machine-like speech that clearly sounds unnatural.

2.3.2 Concatenative Synthesis

Concatenative synthesis was introduced to eliminate the drawback in rule-based synthesis such as the Articulatory and Formant synthesis, which is the difficulty in finding the correct parameter in generating speech (Rashad et al., 2010).

Concatenative synthesis uses the data driven technique by concatenating or joining together different units of recorded human speech made available in an existing speech database to generate speech acoustic waveforms. The pre-recorded human speech units for concatenative synthesis can be in the form of phonemes, words, syllables, diphones and/or triphones. The length of the speech unit usually determines the quality and naturalness of the synthesized speech, where longer speech units are more natural. However, longer speech units require a larger database, which occupies more memory (Rashad et al., 2010). Shorter speech units such as diphone need only small memory size, but it reduces the tendency of synthesizing more natural speech. The two common concatenative synthesis is the diphone and unit selection synthesis.

- **Diphone Synthesis**

Diphone synthesis is the most popular method used for creating a synthetic speech. Diphone refers to speech unit from the middle of one phoneme to the middle of the next phoneme. During the runtime, the target prosody of a sentence is superimposed on these minimal units by means of digital signal processing such as linear predictive coding, Pitch Synchronous Overlap Add (PSOLA), or MBROLA (Pärssinen, 2007). For diphone synthesis, the quality of synthesized speech depends on the phonotactics of the language and the strength of the recorded speech (Indumathi & Chandra, 2012). Diphone synthesis usually suffers from the sonic glitches at concatenation point and the quality of the synthesized speech is generally not as good as the unit selection synthesis but more natural-sounding than the formant synthesis.

- **Unit Selection Synthesis**

A large speech database is used for the Unit Selection synthesis such as the ATR's CHATR (Black & Taylor, 1994). The speech database for Unit Selection Synthesis comes in a variety of units, including phones, diphones, half-phones, syllables, morphemes, words, phrases, and sentences. In unit selection synthesis, the use of digital signal processing is little and therefore, the provision of naturalness is very optimal (Isewon et al., 2014).

The use of small amount of digital signal processing makes unit selection synthesis provides greater naturalness than the concatenative synthesis. An index of the units in the speech database is generated in line with the segmenting or grouping of parameters of the acoustic such as duration, fundamental pitch and places syllable and neighboring phones.

It is very difficult to differentiate between the synthetic speech of the best unit selection synthesis with actual human speech. However, the biggest limitation of the Unit Selection is the very large size of the speech database, making the synthesis time to be very long (Isewon et al., 2014). Another drawback of concatenative synthesis is that it can only synthesize speech units that are available in the database.

2.3.3 Statistical Parametric Speech Synthesis

Statistical Parametric Speech Synthesis (SPSS) based on the Hidden Markov Model (King, 2010) and Deep Neural Network (Zen et al., 2013) is the state-of-the-art speech synthesis technique in recent times. SPSS make use of statistical parameters of speech (spectral and excitation) in the form of an acoustic model. The speech acoustic model was developed from the training process. During the synthesis, these parameters are extracted from the model and converted to speech signals by a vocoder. SPSS produces a fairly natural synthetic speech and flexible voices. SPSS serves as an alternative to overcome the limitation of previous techniques as it only stored parametric representation of sound in speech generation (Rashad et al., 2010).

- **Hidden Markov Model (HMM) Synthesis**

HMM-based synthesis is a type of SPSS that uses symbolic parameters generated from the natural language processing unit to generate speech. HMM-based synthesis has two major advantages over the unit selection synthesis, which are; it is free from audible bugs, and the size of the footprint is very small, unlike the unit selection synthesis that has a very large storage footprint. The smaller size of the storage footprint makes the HMM-based synthesis implementable in hand-held devices. HMM-based synthesis has been developed for many languages such as English, Portuguese, Mandarin, Japanese, Swedish, German, Korean, Slovenian, and so on. (Zen et al., 2009).

Figure 2 shows the Architecture of HMM-based synthesis (Zen et al., 2009), that comprises of two parts, which are training and synthesis. During the training, the recorded human speech parameters (excitation and spectral) together with their phonetic transcription (known as labels) are used to generate the speech acoustic model. During synthesis, the phonetic transcription (labels) of the text to be synthesized was used as the reference to generate the speech waveform from the parameters available in the speech acoustic model.

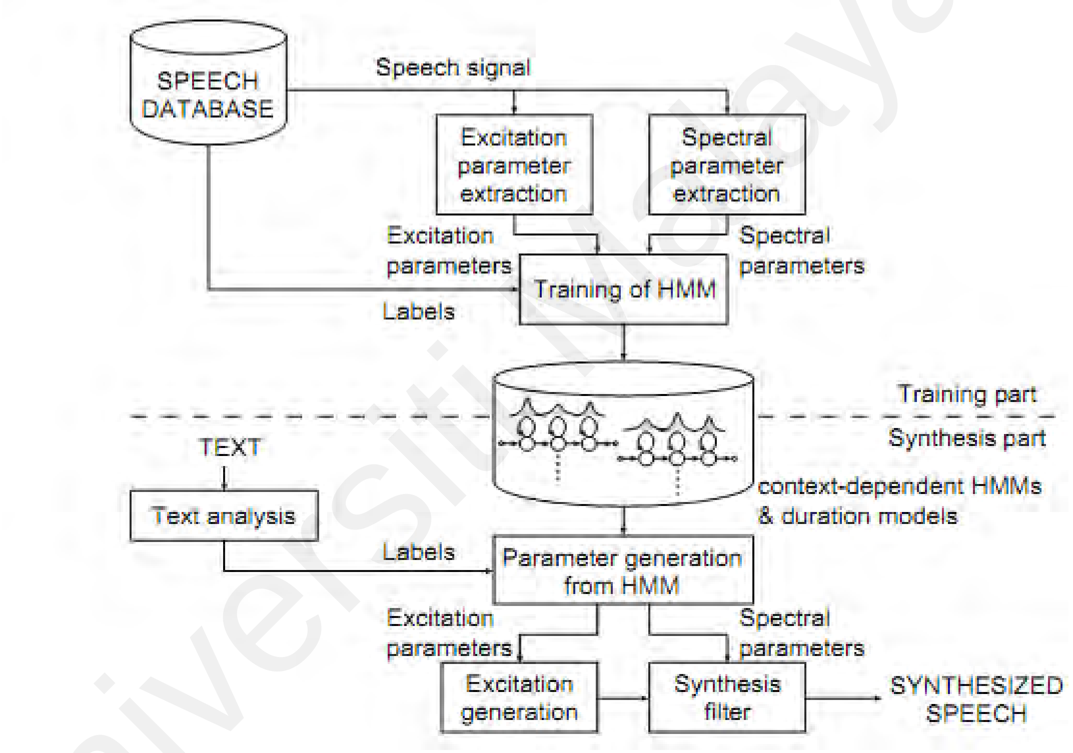


Figure 2.2: Architecture of HMM-based synthesis system (Zen et al., 2009).

The first language to pioneer the development of the HMM-based TTS system was the Japanese (Yoshimura, Tokuda, Masuko, Kobayashi, & Kitamura, 1999). Table 2.1 shows the summary of the HMM-based TTS system developed for various languages. As shown in Table 2.1 HMM-based TTS system has yielded a progressive result in terms of intelligibility and naturalness using small training data with very small runtime engine as compared to the previous Unit Selection TTS system.

Table 2.1: Summary of Speech Synthesis System developed using HMM-based synthesis

Author	**Language**	**Language Status**	**Data Size**	**Training Data**	**Testing Data**	**Performance**
Yoshimura et al., 1999	Japanese	Well-resourced language	Existing database (ATR Japanese speech database)	450 sentences	Not available	The research demonstrate the adaptability of HMM-based in TTS system
Sher, Chiu, Hsu, & Chung, 2010	Taiwanese	Under-resourced Language	Not mentioned	Not Mentioned	Not available	Performance = 4.0 of 5.0 (in terms of naturalness and intelligibility)
Tokuda, Zen, & Black, 2002	English language	Well-resourced language	Existing database (CMU communicator database)	524 sentences	Not mentioned	Very small runtime engine with less than 1MB
Qian, Soong, Chen, & Chu, 2006	Mandarin Chinese	Well-resourced language	1000 sentences	1000 sentences	100 sentences	Using LSP & dynamic features of adjacent LSP produces high quality synthetic speech than conventional and other approach
Hanzlíček, 2010	Czezh	Well-resourced	Existing database	10 minutes, 1 hour and 5 hours speech	96 sentences	The performance of the developed system shows that HMM-based TTS system with STRAIGHT is comparable with Unit Selection TTS system

- **Deep Neural Network (DNN) based speech Synthesis**

DNN-based acoustic models have the potentiality of generating natural sounding synthesized speech by efficiently offering a distributed representation of complex dependencies between acoustic and linguistic features. DNNs are feed-forward artificial neural networks (ANNs), which has three sets of layers; the input layer, the output layer, and the hidden layer (hidden layer may have more than one set of layers). DNN has achieved a remarkable progress in recent years in many machine learning areas including for the development of acoustic models for SPSS using speech parameters such as phonetics, syllabic and grammatical ones. DNN-based TTS system produces more natural speech because the training data is presented by the mapping function of the linguistic features (inputs) with the acoustic features (outputs).

Although DNN acoustic model on deep architectures has better noise robustness and reasonable performance than HMM, DNN suffered a setback on slow in training. However, with the introduction of Mixture Density Networks (MDNs), it overcomes the limitations in DNN-based acoustic modelling for speech synthesis such as absence of variances and the unimodal nature of the objective function (Aroon & Dhonde, 2015).

..

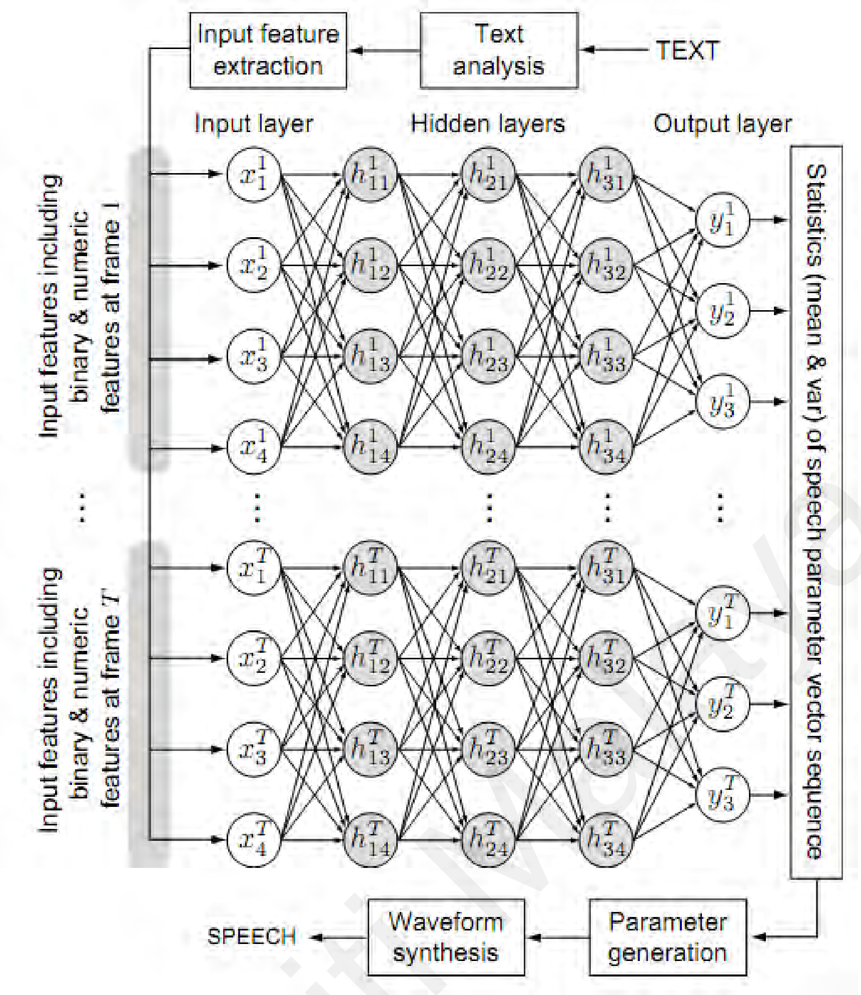

Input feature extraction
Text analysis
TEXT
Input layer
Hidden layers
Output layer
Input features including binary & numeric features at frame 1
Input features including binary & numeric features at frame T
Statistics (mean & var) of speech parameter vector sequence
SPEECH
Waveform synthesis
Parameter generation

Figure 2.3: A speech synthesis framework based on DNN (Zen et al., 2013).

Figure 2.3 above illustrates the framework of DNN-based synthesis. At the beginning, the text input to be synthesized is converted into a sequence of input features { x_n^t}, where x_n^t denotes the *n*-th input feature at frame t. These features include binary answers to questions about the linguistic contexts (e.g is-current-phoneme-aa?) and some numerical data (e.g the number of words in the phrase, the relative position of the current frame in the current phoneme, and duration of the current phoneme).

The input features are then mapped to output features {y_m^t}, by a trained DNN using the forward propagation, where y_m^t denotes the *n*-th output feature at frame t. The output features include the spectral and excitation parameters and their time derivatives (dynamic features). The weights of the DNN can be trained using pairs of input and output features extracted from the training data, the same way as the HMM-based TTS system. In general, DNN-based and HMM-based can share text analysis, speech parameter generation, and waveform synthesis modules. Only the mapping process cannot be shared among the two.

One of the limitations of the decision tree-clustered context-dependent HMM is the inefficiency to express complex context dependencies (Esmeir & Markovitch, 2007). DNN-based TTS system has overcome this limitation. Zen et al (2013) conducted an experiment by developing two speech synthesis systems, which is the HMM-based and the DNN-based using training data consisting of about 33,000 utterances of US English language. A subjective and objective evaluation was conducted that shows the performance of the DNN better than the HMM in terms of addressing the limitations of the conventional decision tree-clustered content-dependent HMM-based approach (Zen et al., 2013).

Rebai and BenAyed (2015) developed a TTS system for Arabic language using DNN with diacritic functionality. The Arabic language is one of the largest spoken languages in the world with more than 400 million speakers. Many HMM-based TTS systems have been developed for the Arabic language but reading Arabic text without diacritic marks is a challenge for those systems. Diacritic marks are used in determining the correct pronunciation of Arabic text and many modern Arabic writers write Arabic text without diacritization. The Arabic language has three different level of lexical stress (i.e primary, secondary and unstressed stress) for each syllable in a word, and these syllables are uttered with different stress, while the last syllable of the word is always unstressed. The identification of the word syllables stress depends on the set of rules. The developed TTS system using DNN has high intelligibility (Rebai & BenAyed, 2015).

Another advantage of SPSS such as HMM and DNN is its ability to change speaker characteristics, speaking style and emotion (Zen et al., 2009). Several studies have shown that the DNN-based TTS system has delivered a promising result. Wu, Swietojanski, Veaux, Renals, and King, (2015) investigated the adaptability of the DNN in speaker adaptation. In their study, a voice bank corpus recorded by about 96 speakers (41 males and 55 females) were used to train the DNN, in which a male and female speakers were used as target speakers. An adaptation data of 10 and 100 utterances were used separately for both target speakers. An objective and subjective evaluation was conducted, where the result confirmed the flexibility of the DNN-based TTS system with better performance than the HMM-based adaptation (Wu et al., 2015).

Having succeeded with the Maximum Likelihood Linear Regression (MLLR) (Leggetter & Woodland, 1995) and Maximum a posteriori (MAP) (Gauvain & Lee, 1994) techniques for speaker adaptation for HMM-based TTS system, recently Wu et al. (2015) conducted a research to investigate the adaptability of speaker voice by DNN. In their research, they have conducted a three-stage transformation. First is the feature space transformation, second is augment speaker-specific features as input to the neural net, and third is the model adaptation. The objective and subjective evaluation show that DNN has better performance than HMM in term of naturalness and speaker similarity (Wu et al., 2015).

Table 2.2 provides the summary of speech synthesis system developed using the DNN-based technique. From Table 2.2, it can be seen that the DNN-based TTS systems made a remarkable progress for many of the well-resourced languages like English due to its better performance than conventional HMM-based TTS system. However, there was no similar development for the under-resourced language.

In this research all comparism was done base on the methods, technique or approaches used for the development of TTS for under-resourced languages, this is to examine which technique yield the best performance for rapid development of TTS system for under-resourced languages. In our review DNN has been shown to have a better performance than HMM for the development of TTS system for well-resourced language especially in reducing the word error rate, this motivated us to experiment its performance for the developments in under-resourced languages.

Table 2.2: Summary of TTS System developed using DNN-based synthesis

Author	**Language**	**Data Size**	**Performance**
Wu & King, 2016	British English (British male professional speaker)	2,400 utterances as training set, 70 utterances as development set and 72 utterances as testing set.	Improves the synthesis performance without much increase in the computational complexity
Wu et al., 2015	English language	Voice Bank corpus (41 male and 55 female speakers)	better performance than the HMM baseline in terms of naturalness and speaker similarity
Narayanan & Wang, 2014	English language	CHiME-2 Corpus.	Performance = 6.7% better than the previous best result

Fan *et al.*, 2016	Mandarin and English language	900 Mandarin & 900 English sentences each for the 3 speakers.	Polyglot system
			Naturalness = 2.44
			Similarity = 2.13
			Monolingual
			Naturalness = 2.69
			Similarity = 2.71

2.3.4 Summary of the Speech Synthesis Techniques

This section summarizes the existing speech synthesis techniques, their merit and demerit. It also summarizes the features of the techniques.

Table 2.3: Various types of Speech Synthesis Techniques

TECHNIQUE	FEATURE	MERIT	DEMERIT
Articulatory Synthesis	Speech is generated by direct simulation of human voice	Produces natural synthesize speech	-Difficult in obtaining data for articulatory model. -Difficult in finding the balance between highly accurate model and easy to design and control
Formant Synthesis (Rule-based synthesis)	Generate speech using set of rules	Good in producing non-nasal and fricatives sound	-Produces machine like speech which is clearly unnatural
Concatenative Synthesis	Generate speech by connecting natural pre-recorded speech units.	Good for synthesizing short units of words	-Only synthesize phones that are define within the speech unit inventory. -Requires large amount of memory space.
Diphone Synthesis		Produces more natural speech than formant synthesizer	-It is discontinuos. -Not good for languages with lot of inconsequence in the pronunciation rules.
Unit Selection Synthesis	Many types of recorded speech units including diphone	-It produces natural sounding speech -It preserve the original voice of the actor	-Requires large amount of database in speech recording. -The process of synthesis is very slow
HMM Synthesis	The HMM synthesis is a parametric synthesis technique,	-Less memory is needed to store the parameters of the model. -More variation are allowable -Easy to be adopted in hand-held devices	-The naturalness may be lower than the best Unit Selection Synthesis

Deep Neural Network (DNN)	Uses specific parameters in the production of speech.	-Noise robustness -High intelligibility and natural of synthetic speech	-Difficult to train -High computational cost

2.4 The Development Process of DNN-based TTS system

This section discusses some of the development process involved for DNN-based TTS system.

2.4.1 Resources required for DNN-based TTS development

HMM-based synthesis share similar resources with DNN-based synthesis except in the mapping module (Zen, Senior, & Schuster, 2013). Many researchers have used the existing resources of HMM-based to develop DNN-based TTS system by only replacing the context dependent hidden Markov model with neural network. Resources like labels, acoustic features are always required.

2.4.1.1 Preparing labels/Transcription

In TTS system conversion of text corresponding to each audio file to labels is always important. Potard, Aylett, Baude, and Motlicek (2016) use Idlak front-end to generate the “full” labels for the training and synthesis of the DNN, which is further converted to numerical values and calculate the output duration using forward propagation. In another research conducted by Lazaridis, Potard, and Garner (2015), they uses a conventional and freely available TTS front-end for labels generation which is further converted to numerical and binary values using Kaldi toolkit. Zen et al. (2013) uses senones as the labels for the DNNs.

2.4.2 Feature Extraction

The first stage to the development of a TTS system is feature extraction. It is an essential stage in which audio analysis is performed on recorded human speeches to collect information such as pitch, power, and vocal tract configuration. The speech signal is then parameterized into sequence of vector, using the spectral analysis technique. Feature extraction is also performed to discard unwanted information such as noises and other irrelevant sound that can distort the TTS system. Feature extraction is an essential step for training process in which raw speech is parameterized into sequence of feature vectors. During this process the entire audio files is converted into MFCC, Filter Bank Energy (FBE) using feature extraction module. This is to reduce the amount of preprocessing required during the training.

2.4.3 Training the Deep Neural Network

Forced alignment is the process of taking the text transcription of an audio speech segment and determines where the time particular words occur in the speech segment. In DNN running a feed-forward algorithm within the HMM framework required likelihood acoustic feature, the forced alignment is performed to fine-tune the DNN acoustic model. This process is important for training labels that are highly unbalance with many silence frames (Hinton et al., 2012). This process requires assigning a senone label to each acoustic output frame in each training utterance. A forced alignment of the ground-truth transcriptions is used to generate a sequence of senone labels for each utterance which is consistent with the word transcription for the utterance.

A DNN is a feed-forward, artificial neural network with multiple hidden layers between its input and output. For each hidden unit *j*, a nonlinear activation function *f(')*, is used to map all inputs from the lower layer, xj ,to a scalar state, yj , which is then fed to the upper layer,

$$Yj = f(xj) \quad (1)$$

Where

$$X_j = b_j + \Sigma_i \, y_i \, w_{ij} \tag{2}$$

and b_j is the bias of unit *j; i* is the unit index of lower layer; w_{ij} is the layer below. Generally the activation function *f(‘)* is chosen to be sigmoid function:

$$f(\, x_j \,) = \frac{1}{1 + e^{-xj}} \tag{3}$$

where the input-output mapping is defined by a logistic regression, or a hyperbolic tangent (or tanh) function:

$$f(\, x_j \,) = \frac{e^{xj} - e^{-xj}}{e^{xj} + e^{-xj}} \tag{4}$$

which is a rescaled version of the sigmoid, and its output range is [-1, 1] instead of [0, 1]. All weight and biases are generally initialized in the pre-training and then trained by optimizing a cost function, which measures the discrepancy between the target vectors and the predicted output with a back propagation procedure. Given a fixed training set $\{x^{(1)}, y^{(1)}, \ldots.. , x^{(T)} , y^{(T)}\}$ with *T* training examples, the cost function to be minimized is defined by

$$\mathrm{C} = \frac{1}{2T} \, \Sigma_{t=1}^{T} \, \| f(\, x^{(t)}) - y^{(t)} \, \|^2 \tag{5}$$

To prevent over-fitting, a regularization term is added into equation 5. Also, learning can be simply terminated when the performance on a held-out validation set starts to deteriorate. The DNN is trained by using the batch gradient descent. It is optimized by a “mini-batch” based stochastic gradient descent algorithm.

$$(W^l ,\cdot b^l) \leftarrow (W^{l\,\cdot} \, b^l) + \varepsilon \, \frac{\partial \mathrm{C}}{\partial (W^l, b^l\,)} \, , \; 0 <= l <= L \tag{6}$$

Where *ε* is a preset learning rate.

The cost function in eq. (5) is often used for classification and regression problems. TTS is a regression problem, the outputs are first scaled to ensure that they lie in the range of [0, 1] or [0, -1] if a tanh activation is used. Nonlinear activation function can allow the neural networks to compute non-trivial problems by using only a small number of nodes. However, for neural network based regression, a nonlinear activation function is normally adopted for the hidden layers, while linear activation function is employed only at the final output layer.

2.5 TTS systems for Under Resourced Languages

In the recent years, many of the under-resourced languages have gained remarkable progress in the development of TTS system. Statistical parametric speech synthesis based on the HMM has greatly contributed to the development of TTS system for the under-resourced language. Building a TTS system from scratch is expensive and requires many resources such as expertise, complex rules for text normalization, labels, and so on. For under-resourced languages, it is often hard to obtain those relevant resources, so there have been several works that aim at identifying suitable techniques for the development of a TTS system for under-resourced language using minimal resources. Table 2.4 provides the summary of existing TTS system developed for under-resourced languages.

HMM-based TTS system offers several advantages, especially for the under-resourced languages, which includes change of speaker voice (Balyan, Agrawal, & Dev, 2013), used of well-resourced language resources for under-resourced TTS-system development (Mumtaz et al., 2011), small storage footprint (Boothalingam et al., 2013), rapid system development (Balyan et al., 2013), and so on. HMM-based TTS system has the ability to synthesize the voices of different speaker, styles, and emotions as well as with the ability for adaptations of speech. Most importantly, HMM allows the use of the existing resources of another language for the rapid development of TTS system for a new language with little resources, which also reduces cost and time for the development of the new system. HMMs allow the use of resources of one language for developing the TTS system for another language that lacks resources like phonetic transcription, segmental labels, contextual factors, and so on through the use of cross-lingual techniques.

The DNN-based TTS system also has an equal ability as the HMM-based TTS system and in some instances performed better than the latter. Despite the benefit of DNN-based TTS system, at the present moment, it was yet to be developed for under-resourced languages.

Table 2.4: Summary of existing TTS systems developed for under-resourced languages

Author	**Language**	**Data Size**	**Training Data**	**Testing Data**	**Performance**
Van Niekerk & Barnard, 2009	Afrikaans IsiZulu Setswana	134, 150 & 332 Utterances	Not mentioned	10 – 20 utterances each	The research shows that baseline HMM segmentation is more convenient, robust, and accurate
Mullah, Pyrtuh, & Singh, 2015	Indian English	1000 utterances		5 sentences	Naturalness: slightly greater than 3.0 out of 5.0 Intelligibility: slightly less than 4.0 out of 5.0
Boothalingam et al., 2013	Tamil	Not mentioned	Not mentioned	Not mentioned	Naturalness & Intelligibility = 3.86 out of 5.0
Ferreira et al., 2016	Mirandese	7 hours speech data	Not mentioned	Not mentioned	The system is relatively intelligible
Justin, Mihelič, & Žibert, 2016	Slovenian	2 min 21sec. speech data	2 min. 21 sec. speech data	30 sentences	The research shows that manual phoneme mapping by expert yield better performance than the automatic phoneme mapping and baseline method
Mukherjee & Mandal, 2014	Bengali	C-DAC corpus	816 sentences	10 sentences	The performance of system is 3.6 out of 5 in terms of intelligibility and naturalness

Gutkin, Ha, Jansche, Kjartansson, et al., 2016	Bangladeshi Bangla	1891 utterances	Not mentioned	100 sentences	Server LSTM-RNN = 3.403 Embedded = 3.519 HMM = 3.43
Mumtaz, Ainon, Roziati, Don, & Gerry, 2011	Malay	1000 sentences	1000 sentences	50 utterances	Intelligibility Male = 99.33 Female = 99.11

2.6 Developing a TTS System for Under-Resourced Languages

This section provides the review on the development of TTS systems for under-resourced languages, such as resource accumulation process, and methods used for the developments of those systems.

Some of the basic resources used for the development of a TTS system for under-resourced languages include speech database, utterances, transcription, and labels.

2.6.1 Recorded Speech Database

A speech database is a collection of sound file and its transcription. The development of speech database is usually expensive and time-consuming, which prevent the resource accumulation for most of the languages unless there is commercial justification for the resources. As such, many researchers accumulate small speech data and uses resources of other well-resourced languages in their research. A good quality speech database for TTS system is the one that is phonetically rich, limited variation of speech, noise free, and limited number of speakers (Mandal & Datta, 2007).

Boothalingam et al. (2013) has developed a HMM-based TTS system for Tamil using 12 hours recorded speech data consisting of 3,732 sentences collected from Tamil novel, news, sports, and so on. The recording was conducted by a female native speaker. Only five minutes speech data (about 40 sentences) were manually segmented (Boothalingam et al., 2013).

Mumtaz et al. (2011) developed a corpus of about 1,000 phonetically rich and balanced sentences for Malay collected from newspapers, textbooks and some reading materials for developing a Malay TTS system. The recorded speech database consists of 2,763 words, 6,599 syllables, and 20,655 phones. The recording which was conducted by a male and female actor is totaling of 4.09 hours speech recording. 50 sentences were used for synthesis and testing.

A Slovenian speech database with the length of 2.21 minutes uttered by one speaker, and about 31 utterances was created for the development of Slovenian TTS System using the automatic cross-language phoneme mapping technique. The database was phonetically rich and balance with about 1,972 labeled phones (Justin et al., 2016).

In the development of the Indian English language TTS system, a speech database containing 1,000 sentences uttered by one female speaker in a studio environment. The database covers all English phonemes of recording duration of about 3 hours and was recorded at 44 kHz, 16-bit mono format which was later converted to 16 KHz sampling rate for the experiment purpose (Mullah et al., 2015)

A recorded speech database comprises of 613 sentences, including 200 phonetically adjusted sentences for the Brazilian Portuguese spoken in Rio de Janeiro was recorded by a male Brazilian Portuguese speaker at a sampling rate of 48 kHz with 16-bits for each sample. The database was posteriorly down sampled to 16 kHz. The phonetic labeling for the recorded speech database was performed in two stages: (1) text-to-phoneme transcription using a phonetic transcriber software; (2) label correction of the transcriptions obtained in step (1) (Maia, Zen, Tokuda, Kitamura, & Resende Jr, 2003). Table 2.5 shows the summary of the existing speech data recorded for the under-resourced languages.

Table 2.5: Summary of the existing recorded speech data of under-resourced languages

Author	**Language (Speaker)**	**Data Size**
Boothalingam et al., 2013	Tamil (female speaker)	3732 sentences recorded 40 sentences for training

Mumtaz et al., 2011	Malay (male and female speaker)	1000 sentences recorded 50 utterances for synthesis
Justin et al., 2016	Slovenian (one speaker)	31 utterances
Mullah et al., 2015	Indian English (female speaker)	1000 sentences recorded
Maia et al., 2003	Brazilian Portuguese	613 sentences recorded

2.6.2 Labels/Utterance file and Transcription

The `labeling' of a recorded utterance involves the temporal definition and naming of units with reference to the acoustic signal. These `units' may be temporarily discrete or overlapping and may be defined in acoustic, physiological, phonetic or higher level linguistic terms. It is clear from this definition that, apart from the orthographic representation, other levels are necessary for this particular domain (Barry & Fourcin, 1992). The purpose of labeling is to enable explicit speech knowledge for the development of SPSS TTS system.

A Brazilian Portuguese HMM-based TTS system was developed, where the labeling was performed in two steps. The first step was text-to-phoneme transcription using a phonetic transcriber software. The second step is the correction of the labels obtained in the first step. The phonetic segmentation (Time label boundaries) was created using the flat-start training HMM available in the HMM Toolkit (HTK), where a time alignment was performed through the Viterbi algorithm for determining time boundaries for each phone. Then, 50 manually corrected sentences were used for re-training the database using the bootstrap method. This method of correcting and posteriorly re-training improves the time label boundaries and thus decreasing the need for manual correction. A total of 80 sentences were segmented (Maia et al., 2003).

For the development of Bengali TTS system, a manual labeling was performed for the prosodic labeling of all the sentences for marking of a phoneme, syllable, and word boundaries along with the appropriate part of speech (POS). The marking of the POS was performed during the generation of the context-dependent label of a given text, and word labeling was performed based on certain Grapheme-to-Phoneme rules. The prosodic word labeling was carried out manually and a total of 816 sentences were used for the training (Mukherjee & Mandal, 2014).

The Festvox toolkit was used for the development of TTS system for Marathi, an Indo-Aryan language. Festvox was used for building new voice including defining the phoneset, tokenization and text normalization, lexicon and grapheme-to-phoneme conversion, and syllabification and stress implementation (Kayte & Gawali, 2015). During the tokenization of Marathi text, a white space and punctuation marks were adopted, while grapheme-to-phoneme algorithms were used for generating a phonemic or phonetic description based on the orthographic linguistic representation developed manually by linguistic experts. Festival syllabification algorithm was used for the production of syllabification and stress (Kayte & Gawali, 2015).

For the HMM-based TTS system for Indian English language, the utterances files were built using corresponding label files. HTK toolkit was used to generate the labels using force alignment technique for generating full contextual labels (Mullah et al., 2015).

Mumtaz et al. (2011) used the time-aligned phone transcriptions to generate the segmental labels for training HMMs of Malay TTS system. Mumtaz et al. (2011) used the existing male and female of US English database to generate the time aligned phone transcription for Malay (Mumtaz et al., 2011). Table 2.6 provides the summary of the labels/utterances and transcription for under-resourced language.

Table 2.6: Summary of labels/utterances and transcription

Title (year)	Labels	Utterance	Transcription
Maia et al., 2003	Using the flat-start trained HMM set	80 sentences	Using phonetic transcriber software
Mukherjee & Mandal, 2014	Manual Method	816 sentences	Grapheme-to-phoneme rules
Kayte & Gawali, 2015	festival syllabification algorithm	Not specified	Grapheme-to-phoneme rules
Mullah et al., 2015	Force alignment approach using HTK	Not specified	Manual segmentation
Mumtaz et al., 2011	time-aligned phone transcriptions using Festival	50 utterances	Grapheme-to-phoneme rules

2.6.3 Speech Acoustic Model

This section discusses the techniques for the development of TTS system for under-resourced languages using the resources of a well-resourced language, particularly for the HMM-based TTS system

Maia et al (2003) developed a TTS system for the Brazilian Portuguese language with a small amount of recorded database of only 200 phonetically balance sentences. A bootstrapping method was used during the training of the database and a text-to-phoneme transcription was adopted using software called phonetic transcriber. The system has the ability to generate speech with acceptable quality (Maia et al., 2003).

Mumtaz et al. (2011) developed a HMM-based TTS system for the Malay language. Malay is a family of Austronesian language spoken in some Asian countries like Malaysia, Indonesia, Singapore, and Brunei. Malay is a phonetic language unlike Mandarin, and it uses Roman alphabet as its writing system. 1,000 phonetically balanced sentences were created by crawling text from different web sources. These 1,000 sentences were uttered by native male and female speakers, with a total recording time of about four hours. A Grapheme-to-Phoneme rule was used in determination of phonemic representation of words during the development of the TTS system.

The Festival speech synthesis system was used to prepare the time-aligned phone transcriptions for Malay using Malay G2P and English CART. The proposed cross-lingual approach simplifies the development of the TTS system for Malay languages by using the existing resources of a well-resourced language. The developed system produces acceptable performance in terms of intelligibility and naturalness despite the use of small database (Mumtaz et al., 2011).

In the effort to reduce the challenges of resources for under-resourced language especially those languages without orthography (i.e standard way of writing), Palkar, Black and Parlikar (2012) proposed a cross-lingual approach as a solution. The Marathi language is assumed as a language with no. The TTS system development for Marathi makes use of the English, Telegu, and Hindi language (sourced language) acoustic model for bootstrapping. A speech database was built for the Marathi (target language) and an Automatic Speech Recognition (ASR) System of the sourced language was applied to generate text from the recorded speech database of Marathi. As such there are three types of speech acoustic model for Marathi. The TTS system of Marathi using Telegu as the source language performed better than the other two language (Palkar et al., 2012). Despite the performance of the developed TTS system, there was an error in generating some of the phones during the bootstrapping process when English is used as the sourced language.

In the effort to identify suitable technique for rapid development of TTS system for under-resourced languages, a TTS system was developed for Bangla language using the HMM-based and Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) (Gutkin, Ha, Jansche, Pipatsrisawat, & Sproat, 2016). In this research, they have proposed the bootstrapping of TTS system for under-resourced language by using crowdsourcing for the development of speech corpus and an existing Hindi text normalization system (Gutkin, Ha, Jansche, Pipatsrisawat, et al., 2016). Three TTS systems were developed, two were based on LSTM-RNN and the other based on the conventional HMM-based TTS system. All TTS systems were able to generate synthetic speech with acceptable intelligibility and naturalness but the conventional HMM outperform the LSTM-RNN technique. In terms of portability in handheld devices, LSTM-RNN outperformed the conventional HMM.

Speech data acquisition is one of the most time-consuming processes during the development of a TTS system for under-resourced languages. Justin et al. (2016) proposed an algorithm that can automatically extract similar acoustic of the target language from the source language. In the proposed solution, an automatic phoneme mapping technique was used where the target language is Slovene language and the English Language is the sourced language.

In many instances, during the development of a TTS system for the under-resourced language, the development involves the creation of a speech database that has to be transcribed and labeled, which is time-consuming and also required experts involvement. However, the use of cross-lingual techniques such as automatically mapping phoneme of the under-resourced language to that of the well-resourced language can be an effective solution when developing the TTS system for under-resourced languages (Justin et al., 2016). Table 2.7 provides the summary of technique used in the development of TTS System for under-resourced languages.

Table 2.7: Summary of techniques used in the development of TTS System for under-resourced languages

Author	Target Language	Resource Language	Techniques	Remark
Maia et al., 2003	Brazilian Portuguese	Portugues e	HMM-based Bootstrapping:Text-to-Phoneme	This research show the effectiveness of text-to-phoneme approach with a little data of the target language.
Mumtaz et al., 2011	Malay	English	HMM-based Cross-Lingual: Grapheme-to-Phoneme	The effectiveness of the cross-lingual technique and G2P rules for bootstrapping.
Palkar et al., 2012	Marathi	-English -Hindi -Telugu	HMM-based Bootstrapping:Speech Recognition	The effect of choosing the source language in bootstrapping TTS system for under-resourced language
Gutkin, Ha, Jansche, Pipatsrisawat, et al., 2016	Bangla	Hindi	HMM and LSTM-RNN Bootstrapping:Manual Transcription	The development of portable TTS system for hand held devices.
Justin et al., 2016	Slovenian	English	HMM-based Cross-lingual:Automatic Phoneme Mapping	A rapid method for the development of TTS system for less-resourced languages but there is deficit in phoneme mapping

2.7 Rapid Development Techniques for TTS system of under-resourced languages

This section discusses some of the techniques used for rapid development of TTS system for Under-resourced languages.

2.7.1 Cross-lingual Adaptation

The collection of a large amount of speech data is cumbersome and time-consuming. Adaptation techniques allow limited speech data from target speaker to be used to develop a TTS system of a different speaker voice. Adaptation consists of three stages, which are, the training, adaptation and synthesis stage as shown in Figure 2.4 below.

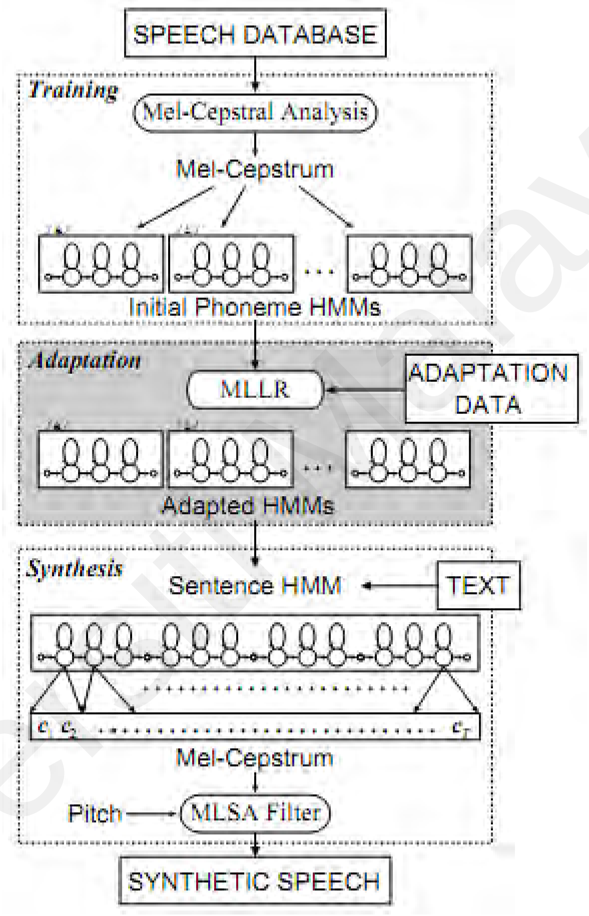


Figure 2.4: Block diagram of speech synthesis system using adaptation technique (Tamura, Masuko, Tokuda, & Kobayashi, 1998).

i Training Stage

During the training stage, spectral and excitation parameters including Mel-frequency cepstral coefficients (MFCCs) and their dynamic features (delta and delta-delta), and the excitation parameter consists of the fundamental frequency (F0) and its dynamic features are extracted. Both the excitation and spectral parameters are modeled by the HMM. The context-dependent penta-phone model is built for each phoneme

State sequence is the observation of how the observable state transforms from one observation to another. In simple term, a state sequence refers to the way the extracted state Mel-cepstrum and log F0 is connected to its neighboring state frame by frame. The training part serves two main purposes, which are the extraction of the parametric representation of speech and the development of speech acoustic model. The input for training HMMs is the recorded speech and HMM-based speech synthesis labels.

ii Adaptation Stage

In the adaptation stage, the given adaptation data is used to calculate the feature vectors, followed by transforming the initial HMMs to the target speaker HMMs by applying the speaker adaptation technique (Tamura et al., 1998).

There are several techniques for speaker adaptation such as the Maximum a Posteriori/Vector field Smoothing (MAP/VFS) and the Maximum Likelihood Linear Regression (MLLR). MLLR outperforms MAP as it uses only one parameter to represent the number of regression matrices, which make it easier to determine the optimum parameters set for voice conversion.

iii Synthesis Stage

In the synthesis phase, text input is transformed into a context-dependent label sequence and then the utterance HMM is constructed by concatenating the context-dependent HMMs according to the label sequence. Next, the sequence of spectral and excitation parameters are generated by the synthetic parameters generation algorithm, this parameters are then concatenated and synthesized speech waveform is produced by a vocoder (Sharma, Adiga, & Prasanna, 2015).

The spectrum part output vector of the HMM is usually based on the Mel-cepstral coefficients together with zeroth coefficients, and their first order derivatives and second order derivatives. In the same way, the state durations of the HMM is modeled by using the multivariate Gaussian distribution. HMM, output vector or Mel-cepstral coefficient controls the synthesis filter during the speech synthesis, which allows speech to be re-synthesized directly by using Mel Log Spectrum Approximation (MLSA) filter. Fundamental frequency F0 and the observation sequence can also be modeled by the HMM, which contain one-dimensional continues values and discrete symbol.

A cross-lingual adaptation is a solution used to speed-up the TTS system development for under-resourced languages, where resources like contextual factors, duration model, and segmental labels are applied from well-resourced languages (Mumtaz et al., 2011).

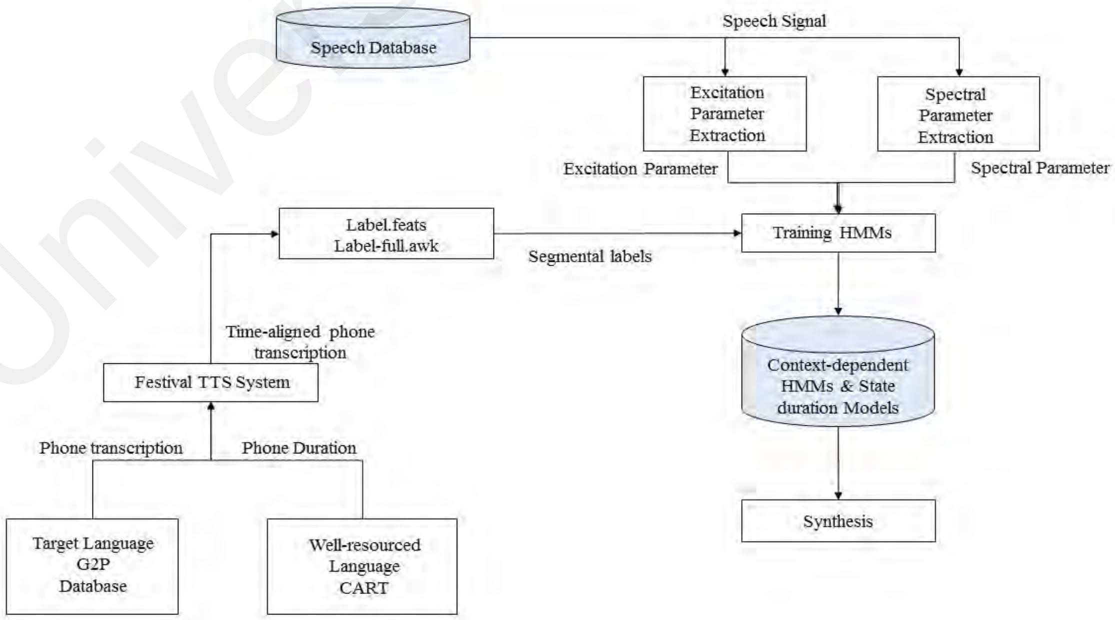

Speech Database
Speech Signal
Excitation Parameter Extraction
Spectral Parameter Extraction
Excitation Parameter
Spectral Parameter
Label.feats
Label-full.awk
Segmental labels
Training HMMs
Time-aligned phone transcription
Festival TTS System
Context-dependent HMMs & State duration Models
Phone transcription
Phone Duration
Synthesis
Target Language G2P Database
Well-resourced Language CART

Figure 2.5: A Cross-lingual approach for development of TTS for under-resourced language Adapted from (Mumtaz et al., 2011).

In figure 2.5, a cross-lingual technique is proposed by Mumtaz et al. (2011) for developing a Malay TTS system using the English resources to generate a time aligned phone transcriptions. This time aligned phone transcription was used by the HMM-based TTS system to generate segmental labels for training HMMs.

When a phone transcription of the target language and phone duration of the sourced language is fed into the Festival TTS system, a time aligned phone transcription is created by comparing the similarities between the target language and the sourced language phonemes, syllabification rules and grammatical classes (like POS). This allows the formation of labels for the target language that are used for the training of HMMs. Mumtaz et al. (2011) has built a G2P database for Malay using English as the basis. However, G2P rules might not be applicable for some foreign words or language that is not phonetic based like Bangla (Gutkin, Ha, Jansche, Pipatsrisawat, et al., 2016).

iv Grapheme-to-Phoneme (G2P) rule

The text is converted into a sequence of phonemes or phone using Grapheme-to-Phoneme rules. Normally text is considered on the word by word basis, thus, languages, where the words are not segmented, can still be processed. Grapheme-to-Phoneme rules enable the use of smaller large pronunciation dictionary. On top of that G2P rules can be used to generate the pronunciation of the words not found in the database. However, rules may be difficult to be coded for languages with very high regular spelling such as Portuguese, and Spanish. Some of the G2P rules techniques (sometimes called Latter-to-Sound rule LTS) are stem and apex rule, rule chain, and finite state transducer.

v Phonetics segmentation and labeling

The HMM-based TTS system phonetic segmentation and labeling requires a number of phonetic and their linguistic information such as phone duration, grapheme-to-phoneme and part of speech tagging for each phoneme. This process can be performed manually but it is expensive and time-consuming. As such, phonetic segmentation and labeling are usually performed automatically using segmentation tools provided in the HMM-based toolkit (Mustafa, Don, Ainon, Zainuddin, & Knowles, 2014).

2.7.2 Bootstrapping

Bootstrapping process is a process used for languages that do not have orthography (i.e. they don't have standard writing form). It is used to speed-up the development of TTS system by creating a phonetic for the target language using the acoustic model of another language through the Automatic Speech Recognition (ASR). The phonetics generated from the ASR system is then used to train the acoustic model for the target language.

Sitaram et al. (2013) proposed a bootstrapping called bootstrapping phonetic transcription, in which a speech corpus of a well-resourced language was used to create a phonetic transcription of the target language, which was then used to develop a TTS system using the phonetics of the ASR system as shown in Figure 2.6 (Sitaram, Palkar, Chen, Parlikar, & Black, 2013).

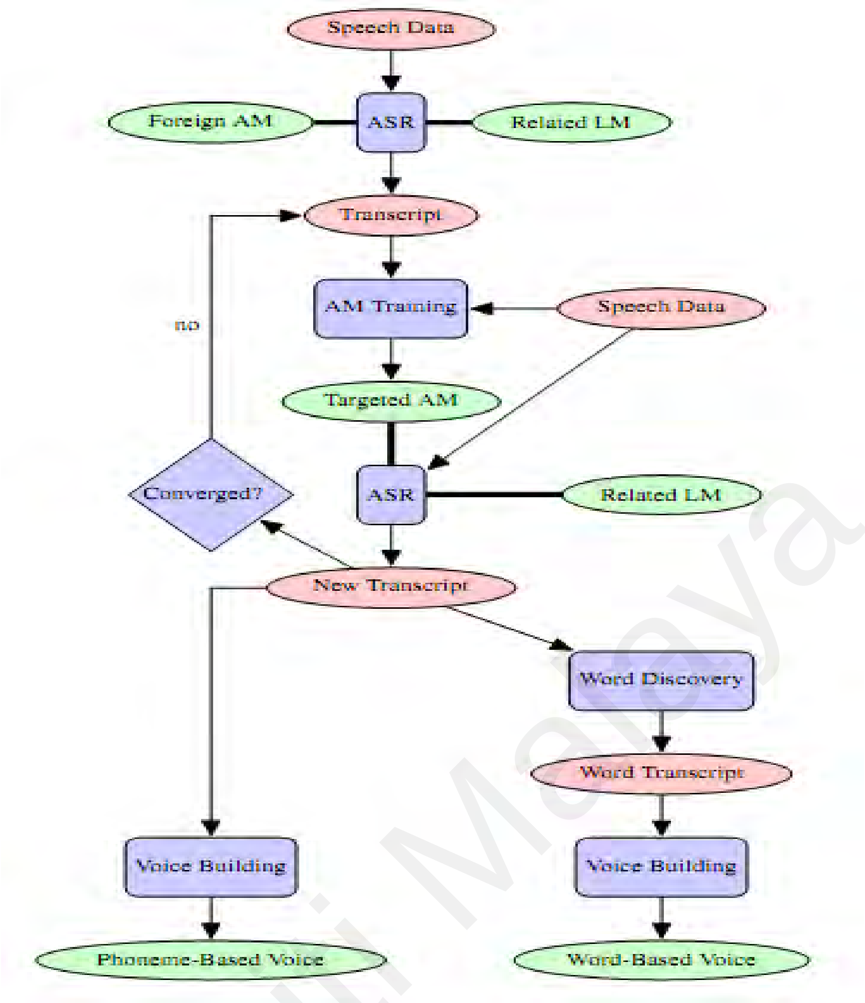


Figure 2.6: Bootstrapping Phonetic Transcription Technique (Sitaram et al., 2013).

2.8 Evaluation of the Speech Synthesis System

One of the main techniques for evaluating the performance of a TTS system is to use human listeners to listen to the synthetic speeches and respond to specific questions objectively or subjectively.

2.8.1 Evaluating the Intelligibility

To evaluate the intelligibility of a TTS system requires the human listener to correctly comprehend the synthesized speech that they have listened. The intelligibility of a TTS system is usually expressed as a percentage of words, sentences or phonemes correctly identified by a group of listeners. Listeners normally listened to the synthesized speech (paragraph, or sentences), and response to a series of questions related to the text content, which are open response questions (subjective) or closed response questions (objective). Open response questions are known to be more sensitive, and therefore are more likely to highlight the differences between the test conditions.

Listeners, however, can make use of syntax and semantic information to help them predict the intended message, even if the synthesized speech itself is not intelligible. Listener's ability for making smart guesses must be accounted for when designing the intelligibility evaluation task. Some of the common methods for intelligibility evaluation are, Diagnostic Rhyme Test (DRT) (Voiers, 1977), Modified Rhyme Test (MRT) (House, Williams, Hecker, & Kryter, 1965), and Semantically Unpredictable Sentences (Stan, Yamagishi, King, & Aylett, 2011).

Yamagishi et al. (2009) uses semantically unpredictable sentences to evaluate the intelligibility of the TTS system (Yamagishi et al., 2009). In Mumtaz et al. (2011), a listening test was used to evaluate the intelligibility of the synthetic speech, using 36 native Malay listeners. In the development of speech synthesis system for Tamil, a subjective evaluation method was conducted using the Mean Opinion Score (MOS) to evaluate the intelligibility of the TTS system. 14 listeners were used, in which only 10 of them were native (Boothalingam et al., 2013). The intelligibility of a Bangla TTS system was also evaluated using the Mean Opinion Score (MOS) (Gutkin, Ha, Jansche, Pipatsrisawat, et al., 2016).

2.8.2 Evaluating the Naturalness

For evaluating the naturalness of synthetic speech, the most widely used evaluation method is the Mean Opinion Score (MOS). The MOS generally includes a 5 point Likert-like scale for ranking the naturalness of the synthetic speech, and the evaluator's task is to evaluate the synthesized speech they have heard using this scale.

Alternatively, naturalness can be evaluated using the Paired Comparison, where listeners are presented the same utterances from two different synthesis systems/techniques sequentially, and evaluators then choose which of the two synthesis speech is of higher naturalness (Yamagishi, Ling, & King, 2008).

Many of the TTS systems developed uses a different form of naturalness evaluation. In Blizzard Challenge 2007, various systems developed by the participating teams use listening tests to evaluate the naturalness of the developed system in which a comparison was performed between the synthetic speech and speech of the original speaker (Yamagishi et al., 2009). In the development of a Bengali TTS system, a listening evaluation was conducted to evaluate the naturalness of the synthesized speech using 5 subjects (3 males and 2 females) (Mukherjee & Mandal, 2014). A Mean Opinion Score was used to evaluate the naturalness of the output (voice) of the Yoruba TTS system using 10 native speakers (Aoga, Dagba, & Fanou, 2016).

2.9 Chapter Summary

The findings of the literature review show that there are two state-of-the-art speech synthesis techniques, the HMM and DNN, where the latter overshadows the performance of the conventional HMM due in term of intelligibility and naturalness. The review also shows that despite the performance of the DNN, it was not developed for any of the under-resourced languages. This review also shows that the resources of well-resourced languages can be used for the development of a TTS system for under-resourced languages. Table 2.8 provides the summary for the development of TTS system for under-resourced languages. This table describe the development of TTS system for under-resourced language, the well-resourced language they use, the technique, method and number of speaker they used for the development. It also describe the data size, utterances, transcription and evaluation method for the development of this systems.

Table 2.8: Summary for Development of TTS system for Under-resourced Languages

Author	Target Lang/ Speaker	Resource Lang/ Speaker	Technique	Method	Speaker	Data size	Utterances	Labels/Transcription	Evaluation
Maia et al., 2003	Brazilian Portuguese	Not stated	HMM	Bootstrapping Text-to-Phoneme	One male speaker	613 Sentences	80 Utterances	Flat-start trained HMM set/ phonetic transcriber software	Not stated
Mumtaz et al., 2011	Malay	English	HMM	Cross-Lingual	One male & one female	1000 Sentences	50 Utterances	time-aligned phone transcriptions using Festival/ Grapheme-to-phoneme rules	Objective Evaluation
Palkar et al., 2012	-Marathi -Hindi -Telugu	-English -Hindi -Telugu	HMM	Bootstrapping/ Speech recognition	Single speaker each	Mar. 30m Hin. 1hr Tel. 1hr	Nil	Transcription generated by ASR	Objective & Subjective Evaluation
Boothalingam et al., 2013	Tamil	Existing Database	HMM	Not stated	One female	3732 Sentences	40 Utterances	Forced-Viterbi alignment	Subjective Evaluation
Mukherjee & Mandal, 2014	Bengali	Not stated	HMM	Not stated	Not state	816 Sentences	80 Utterances	Manual method/ Grapheme-to-phoneme rules	Subjective Evaluation
Mullah et al., 2015	Indian English Language	Not stated	HMM	Not stated	1male speaker	1000 Sentences	5 minutes speech data	Force alignment approach using HTK/Manual segmentation	Subjective Evaluation

Justin et al., 2016	Slovenian	English	HMM	Cross-lingual/automatic phoneme mapping	One speaker	Not mentioned	31 Utterances	Automatically generated from well-resourced language	Subjective Evaluation (AB)
Gutkin, Ha, Jansche, Pipatsrisawat, et al., 2016	Bangla	Hindi	HMM/LSTM-RNN	Bootstrapping	Multiple speaker	1891 Utterances	Not mentioned	Manually transcribe by linguistic	Subjective Evaluation

CHAPTER 3: METHODOLOGY

This chapter describes in details the research methodology adopted in order to achieve the objectives of this research which is to develop a TTS system for Hausa using suitable state-of-the-art techniques with limited resources that can synthesize the input text with an acceptable degree of intelligibility and naturalness. The four main stages in this research include: (1) Problem identification and solution (2) Resource accumulation and preparation (3) development (4) evaluation and results. Figure 3.1 depicts the methodological flow of this research.

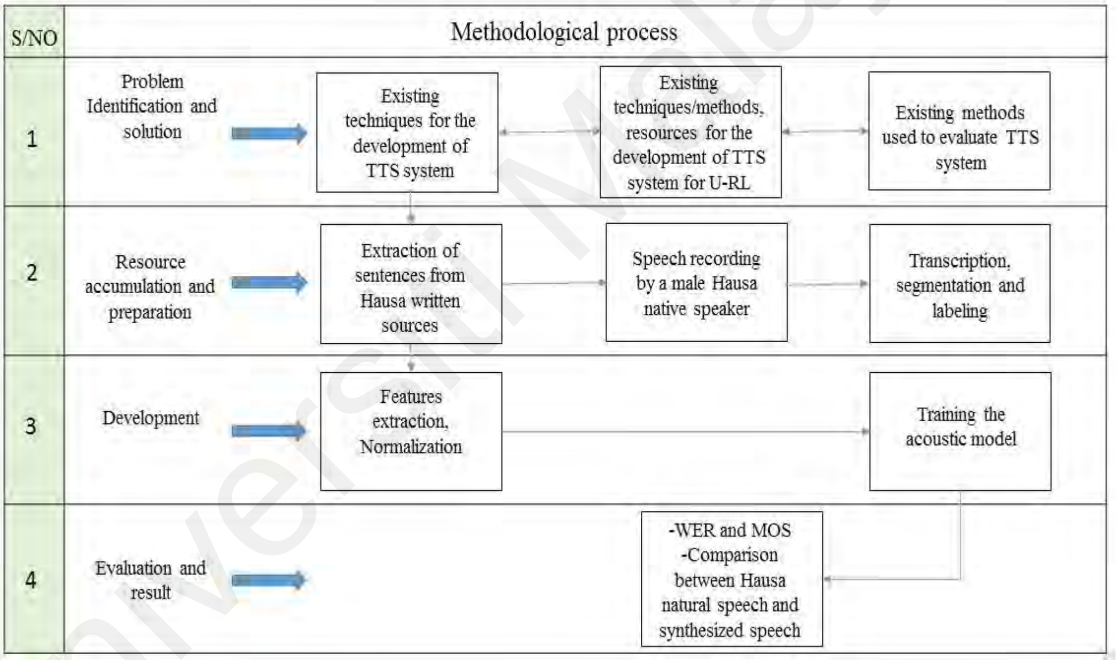

S/NO
Methodological process
1
Problem Identification and solution
Existing techniques for the development of TTS system
Existing techniques/methods, resources for the development of TTS system for U-RL
Existing methods used to evaluate TTS system
2
Resource accumulation and preparation
Extraction of sentences from Hausa written sources
Speech recording by a male Hausa native speaker
Transcription, segmentation and labeling
3
Development
Features extraction, Normalization
Training the acoustic model
4
Evaluation and result
-WER and MOS
-Comparison between Hausa natural speech and synthesized speech

Figure 3.1: Methodological flow of this research

3.1 Problem identification and solution

For identifying the research problem and potential solution, review of the literature was conducted, which consists of four major stages: (i) to investigate the progress of speech technology development for Hausa language and the existing resources for Hausa. (ii) to identify the existing techniques for the development of TTS system, their features, advantages and disadvantages. (iii) To identify the existing TTS systems developed using the Statistical Parametric Speech Synthesis (SPSS) techniques. (iv) To identify the techniques used for the development of TTS system for under-resourced language, the resource accumulation method, the development process and the evaluation method used for the development of an under-resourced TTS system.

3.1.1 Problem Identification

The major problems for the development of a TTS system for the under-resourced languages include the lack of transcribed speech data, pronunciation dictionary and speech acoustic model, and the time and cost in acquiring the components for the development of a TTS system for the under-resourced language. The earlier speech synthesis techniques (such as formant, diphone synthesis etc.) have not been used for the development of TTS system for under-resourced languages because they require many different resources (such as labels, transcriptions etc), which are not available for under-resourced languages and they also require longer time for the development.

While HMM-based has dominated the development of TTS system for under-resourced languages, it suffers some setback from the decision tree used for the training of the speech acoustic model, which resulted in the reduction in the intelligibility of the TTS system.

3.1.2 Possible Solutions

Researchers have proposed different techniques for the development of the TTS system for the under-resourced languages such as the cross-lingual adaptation, phoneme mapping, bootstrapping, and so on. Cross-lingual adaptation makes use of the small resources of the under-resourced language such as text and speech transcription, segmentation and labeling, together with the resources of a well-resourced language.

Phoneme mapping technique supports the development of a new TTS system for under-resourced language with only small amount of recorded speech data from the target language although there are some phonemes that are unmapped due to the non-availability in the resourced language (Justin et al., 2016), as well as dealing with more than one phoneme of the target language. On the other hand, bootstrapping technique support languages without the standard written format (orthography) such alphabetic, syllabic and logographic (example Tibetan language) (Sitaram et al., 2013), providing rapid development of the TTS system, as well as speech-to-speech translation system (Palkar et al., 2012).

A study conducted by Palkar et al. (2012), show that phoneme mapping yield a better result in the development of a TTS system for under-resourced languages provided that the well-resourced language is suited for the target language. However, the selection of the source language is not trivial as it is difficult to find two languages with same phonemes, therefore leaving some phonemes to be unmapped. In Justin, Mihelič, and Žibert (2016), it was found that the phoneme of the target language does not have the exact match with the source language. This situation arises as the target language has more phonemes than the source language or some of the phonemes of the target language is not found in the source language.

Mumtaz et al. (2011) have conducted a study on cross-lingual adaptation that shows a promising result with the high performance in terms of naturalness and intelligibility, which is a suitable technique for less resourced language. The proposed solution in Mumtaz et al. (2011) has solved the problem of previous techniques by manually incorporating phonemes that are not available in the source language.

Bootstrapping technique that commonly used for language without orthography (Palkar, Black, & Parlikar, 2012) & (Sitaram, Palkar, Chen, Parlikar, & Black, 2013) has also produced a good result in the production of synthetic speech and is effective for the development of speech-to-speech system but may not be suitable for TTS system because the speakers of the target language need to learn the orthography of the source language.

More recently, there has been encouraging achievement by the Deep Neural Network (DNN) in speech recognition (Hinton et al., 2012). As such, extensive works have been carried out to investigate the use of DNN for TTS system (Ling et al., 2015). DNN is used in the parametric speech synthesis to replace the GMMs (Gaussian Mixture Models) associated with the leaf nodes of the decision tree. The effectiveness of DNN has overshadowed the conventional HMM-based speech synthesis due to its ability in production of high natural synthesized speech, and many DNN systems were developed for well-resourced languages, yielding a better result than the best HMM-based synthesis, especially in reducing the word error rate by 16.7%

Adaptation technique (Tamura, Masuko, Tokuda, & Kobayashi, 1998) is one of the techniques applied in the HMM-based speech synthesis to change the speaker characteristic, speaker style and emotion (Zen, Tokuda, & Black, 2009). It was also found that the adaptation technique using DNN yield a better result. Despite the progress of DNN-based speech synthesis, it has not been used for the development of TTS system for the under-resourced language.

In this research, we have choose to use DNN and cross-lingual adaptation technique because of yielding performance of DNN in the development of TTS system for well-resourced language and the yielding performance of cross-lingual adaptation for the development of under-resourced language TTS system.

3.2 Resource Accumulation

This section discusses the resources required for the development of a TTS system for the new language, which includes text corpus, speech corpus, transcription, pronunciation dictionary, and the acoustic model.

3.2.1 Text Corpus

The text corpus is one of the important components for language modeling and for the development of the acoustic model for speech synthesis system. It is a collection of a large and structured set of texts. Building a large text corpus can be done automatically using applications that crawled and extract text from web pages. Many of the under-resourced languages acquire text from the Internet for fast text corpora development (Le & Besacier, 2009). The texts crawled such as the Rapid Language Adaptation Toolkit (RLAT) requires several steps such as: (i) removing special characters and empty lines, (ii) removing HTML tags and codes, (iii) identifying and removing other words, phrases and sentences from other language and so on (Schlippe et al., 2012). In this research, the automatic crawler will be used for the creation of text corpus for fast phonetically rich and balanced corpus development.

3.2.2 Speech Corpus

Speech corpus is a database of recorded speech audio files together with the text transcription. For the development of TTS system, a suitable speech database must be identified. Schlippe, Djomgang, Vu, Ochs, and Schultz (2012) have developed a Hausa speech database for the Globalphone corpus, but is not available for research-use and not suitable for TTS system development due to the high variation from many different speakers (about 102 speaker use for the recording). As such, in this research, a speech database is built for the Hausa language, where the speech produced by a Hausa native speaker is used for the development of the speech database.

The database to be developed in this research is a small size database of about 1,046 words (179 sentences). This is because the complexity of building speech database for new languages such as high development cost, time constrains, and the difficulty in obtaining professional speaker, plus the fact that this research will be using English resources for cross-lingual adaptation.

3.2.3 Segmentation and labelling of recorded speech

Segmentation is the process of identifying the phoneme boundary of a speech utterance, and labeling is the process of providing the lexicon and phonological information of the recorded speech. Speech labeling provides the linguistic representations including the phonology and the prosody-related to a recorded utterance. While some resources such as a speech database may be relatively easy to be developed, segmentation and labeling are much more challenging, not only for under-resourced languages but also for well-resourced languages as well. Many of the under-resourced languages including Hausa do not have the existing speech acoustic model for performing the automatic segmentation. Therefore, in this research, there is need to perform the manual segmentation and labeling, which will be performed by the linguistic expert to ensure proper segmentation. In this research, the manual segmentation and labeling were performed using Praat software (Praat is a standalone application that allows sound manipulation and visualization). Using Praat, the recorded speech is segmented and labeled using grapheme-to-phoneme rule, which is similar to the HMM-based speech synthesis.

3.2.4 Pronunciation Dictionary

Although there is an existing pronunciation dictionary for Hausa in the domain of continuous speech recognition, it is not accessible for research purpose. Therefore, there is the need to develop one for this research. A pronunciation dictionary is going to be built using the Grapheme-to-Phoneme G2P method, that is by listing all the phonemes and words of Hausa according to the TIMIT notation which is the standard method used by the existing researches.

3.2.5 Speech Acoustic Model

Speech acoustic model is obtained from the parameterized speech signal into the feature vector that is matched with reference to the acoustic model. The acoustic model is a statistical representation of the sound that makes up each word. In order to match feature vector, compiling the speech sound together with the transcription into statistical representations is required. However, the presence of the external noises makes the acoustic model less accurate and improving the accuracy (in order to achieve robust synthesis) is one of the goals of the researcher model. One way to achieve robustness is using suitable model for feature extraction such as the DNN, which offers more robustness than the conventional HMM.

3.3 Proposed Technique for the Development of TTS system for Hausa

This section discusses the techniques for the development of the speech acoustic model for an under-resourced language.

Researchers have proposed different speech synthesis technique for building the acoustic model for under-resourced languages, but the most common one is the HMM-based synthesis system. In recent time, DNN-based synthesis system was found to offer more intelligible and natural synthesized speech, but it was not developed for any of the under-resourced languages.

A DNN is a feed-forward artificial neural network with multiple hidden layers between the input and output layer, creating a mapping function between the input (i.e. linguistic features) vector and the output (i.e. acoustic features) vector. The speaker adaption using

DNN offer higher performance in terms of speaker similarities than HMM (Wu et al., 2015). The work in Wu et al. (2015) motivates the use of DNN in this research for the development of TTS system for Hausa using the cross-lingual adaptation technique.

Zen et al. (2013) highlighted three major factors that affect the quality of synthesized speech, which is, vocoding, the accuracy of the acoustic model and, the effect of over-smoothing, which DNN have addressed for generating accurate speech acoustic model. Price et al. (2016) have shown that the DNN have reduced the WER by 16.7% as compared to the conventional HMM (Price, Iso, & Shinoda, 2016).

3.3.1 Cross-lingual Adaptation Technique

From the literature review, it has shown that the cross-lingual adaptation technique performed better than the other methods for the development of the acoustic model for under-resourced languages. Cross-lingual adaptation is selected in this research as it is more flexible to be used, as it allows manual incorporation of the target language phonemes that are not found in the source language (Mumtaz et al., 2011).

This research proposed the development of a DNN-based TTS system for Hausa using the cross-lingual adaptation technique. The proposed development includes the accumulation of resources, development of speech database using small-sized Hausa speech corpus (about 1,046 words), and the development of the acoustic model using cross-lingual adaptation technique. The cross-lingual adaptation will make use of English phone duration to generate a time-align phone transcription. English language resources were applied in this research as it has been widely used in previous researches, and English has many similarities with Hausa especially the segmental phonemes (Malah & Rashid, 2015). Table 3.2 shows the comparison of the Hausa and English vowels, and table 3.3 shows the Hausa and English consonant.

Table 3.1: Hausa and English vowels

	Hausa	English
Pure vowels	/i:/, /i/, /e:/, /e/, /Ɔ:/, /ɒ/, /a:/, /a/, /u:/, /ʊ/	/i:/, /i/, /e/, /Ɔ:/, /ɒ/, /a:/, /u:/, /ʊ/, /æ/, /Ʌ/, /ə/, /ə:/
Diphthongs	/ai/, /au/, /ui/	/ei/, /ai/, /au/, /Ɔi/, /iə/, /eə/, /əu/

Table 3.2: Hausa and English consonants

S/No	Description	Hausa sound	English sound	S/No	Description	Hausa sound	English sound
1	Bilabial plosive	B	p, b	17	Alveolar nasal	n	n
2	Alveolar plosive	t, d	t, d	18	Palatal nasal	ɲ	
3	Velar plosive	k, g	k, g	19	Velar nasal	ŋ	Ŋ
4	Labialized velar plosive	kw, gw		20	Bilabial fricative	ɸ	
5	Palatalized velar plosive	kj, gj		21	Palatalized bilabial fricative	ɸj	ŋʒ
6	Glottal plosive	ʔ	[ʔ]	22	Alveolar fricative	s, z	s, z
7	Palatalized glottal	ʔj		23	Post-alveolar fricative	ʃ	ʃ, ʒ
8	Bilabial implosive	ɓ		24	Glottal fricative	h	h
9	Retroflex implosive	ɗ		25	Post-alveolar affricate	tʃ, dʒ	tʃ, dʒ

10	Alveolar ejective	s’		26	Alveolar lateral	l	L
11	Velar ejective	k’		27	Alveolar trill	r	
12	Labialized velar ejective	k’w		28	Retroflex flap	ɽ	
13	Palatalized velar ejective	k’j		29	Palatal approximant	j	j
14	Bilabial nasal	M	M	30	Labio-velar approximant	w	w
15	Labio-dental fricative		f, v	31	Dental fricative		ð, Ɵ
16	Post-alveolar approximant		r	32			

The pronunciation dictionary of English is used in this research as it contains the stress information, which is similar to Hausa. Table 3.1 shows the similarities between Hausa and English TIMIT notation.

Table 3.3: Similarities between Hausa and English TIMIT notation

Phonemes types	**Hausa phonemes in TIMIT notation**	**English phonemes in TIMIT notation**
Vowels	aa, ae, ah, ax, ih, iy, uh, uw, oh	aa, ae, ah, ao, ax, eh, ih, iy, ow, uh, uw
Diphthongs	aw, ay	aw, ay, oy, ey, er,
Nasal	m, n	m, n, ng

Plosive	b, ɓ, f, fy, t, ts, d, ɗ, k, ƙ, g	p, b, t, th, d, dh, k, g
Fricative	f, h, s, sh, z, kh	f, v, s, sh, z, zh, kh, hh
Affricative	c, j	ch, jh
Approximant	y, w, r	y, w, r
Lateral	L	l

3.3.2 Proposed framework for Hausa TTS Development

Figure 3.2 depicts the block diagram of a DNN-based speech synthesis system (adapted from Fan et al., 2016) that consists of training and synthesis part. The training is for the purpose of generating the DNN output. During the training, the acoustic features are extracted from the speech signal with the feature extraction module and the linguistic features, which is going to be generated by the proposed cross-lingual approach. The parameters of DNN are trained by using pairs of input and output features with a mini-batched, back-propagation algorithm.

During synthesis, the text is first analyzed and labels are generated, which is then mapped onto the acoustic features by the trained DNN. For generating smooth parameter trajectories, the dynamic features are used as constraints in speech parameter generation, where predicted features are used as mean vectors and global variances of the training data are adopted for generating speech parameters by maximizing the probability. Finally, the speech waveform is synthesized from the generated parameters with a vocoder.

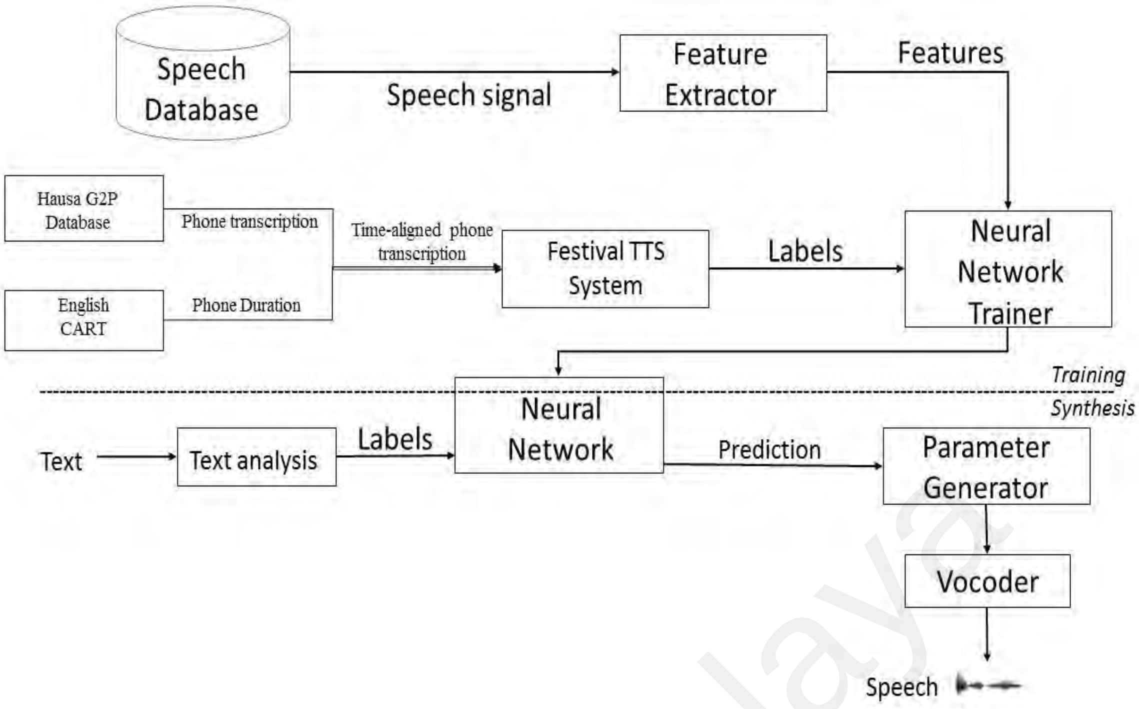


Figure 3.2: The proposed framework for the Hausa TTS system using DNN-based cross-lingual technique. Adapted from (Fan, Qian, Soong, & He, 2016).

3.4 Evaluation

The most effective method for assessing TTS system is by conducting listening tests by the potential native listeners. The method of evaluation used in this work is the most closely related to user acceptance test used in software engineering. In speech synthesis system, the more common attributes that are considered for evaluation are the intelligibility and the naturalness, which will be measured through the listening test.

The evaluation of the performance (i.e the intelligibility and the naturalness) of the developed TTS system for Hausa will make use of the mean opinion score (MOS) (Yamagishi et al., 2009) and word error rate (WER) (Sharma & Prasanna, 2016). The evaluation process will involve the Hausa native listeners.

3.4.1 Naturalness

The naturalness is the measures of the degree of similarities between the human speech and the synthesized speech. Listeners usually listen to the synthetized speech and compare the synthetized speech with the actual human speech and measures how close these two speeches are. In this research, the naturalness is measured using the mean opinion score (Weiss et al., 2007). A 5 points Likert-scale is used, where the listeners will rank the synthetic speech from 1(Highly unnatural) to 5 (Highly natural). The 5 point Likert scale is the most widely used and simple method for the evaluation of the naturalness of the system (Clark, Podsiadlo, Fraser, Mayo, & King, 2007). The sample of the questionnaire for the listeners is presented in Appendix A.

3.4.2 Intelligibility

Intelligibility is the ability of human listeners to correctly comprehend the synthesized speech. For intelligibility test, the listeners will listen to the synthesized speech and write what they listened (please see Appendix A for the sample of the questionnaire). Intelligibility is measured using the Word Error Rate (WER) (Fraser & King, 2007), which is calculated as follows:

$$\text{Word Error Rate} = \frac{S+D+I}{N}\, X\ 100 \qquad (1)$$

Where N is the total number of utterances, S is the number of substitution error, D is the number of deletion error and I is the number of insertion error.

3.5 Chapter Summary

The aim of this research is to accumulate the resources of Hausa, an under-resourced language and identify the suitable techniques that can be used to develop a TTS system for this language with acceptable naturalness and intelligibility. Unfortunately, from the review of the literature, it was discovered that the major problem for under-resourced language is the lack of the resources needed for the development of TTS system. However, it has been discovered that the HMM-based TTS has been developed for many of the under-resourced languages due to its ability to adapt the resources of a well-resourced language. However, no similar work was carried out using DNN for under-resourced language although DNN performs better than the HMM.

This research proposed the development of a TTS system for Hausa language using the cross-lingual adaptation technique using the English resources. However, instead of using the conventional HMM for speech acoustic model development, this research uses the DNN to improve the intelligibility and the naturalness.

CHAPTER 4: RESOURCE ACCUMULATION AND PREPARATION

This chapter explains in detail the resource accumulation for the development of the TTS system for Hausa, which includes text corpus development, speech corpus development, transcription, and segmentation and labelling.

4.1 Structure of Hausa Language

Hausa is the biggest dialect of the Chadic family in sub-Saharan Africa. It is spoken by the masses from the northern Nigeria and the Southern Niger Republic. Hausa has the biggest number of local speakers of any Chadic dialect, which is widely used for exchange and business in numerous part of Africa (Adamu, 1984).

Hausa utilizes tones to show the syntactic structures and the presence of auxiliary explanations made it to its numerous consonants. The verbal framework is astounding for having preverbal buildings, arched for individual and tense-viewpoint, and going before constant stems. Hausa sentences have the typically Chadic subject-object-verb word order. Despite the fact that Hausa is not the national language of Nigeria and Niger, it ruled the elementary schools instructive dialect and broadly utilized as a part of the media and has a broad writing, being likewise utilized as a part of government and business (Adamu, 1978).

- **Hausa Phonology**

The Hausa language has three syllable structure, which is the consonant-vowel (CV), consonant-vowel-vowel (CVV) the vowel-vowel (VV) or diphthongs, and consonant-vowel-consonant (CVC). The vowels have 10 monophthongs and two diphthongs, the monophthongs consist of short and long sound (Sani, 1999). Table 4.1 shows the Hausa language monophthongs.

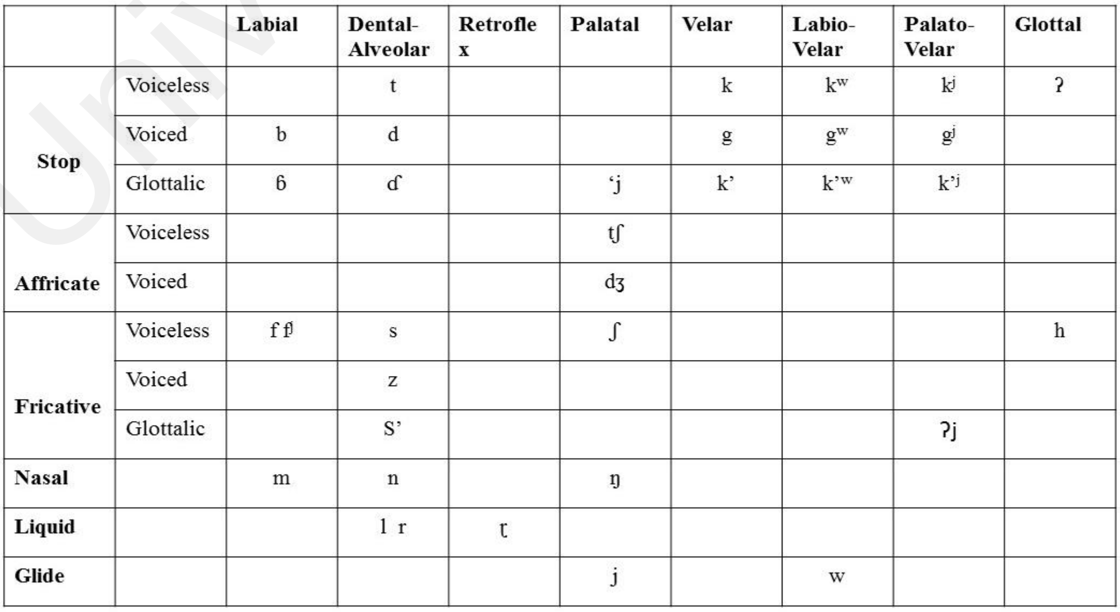
Table 4.1: Hausa language monophthongs

	Front	Central	Back
High	i i:		u u:
Mid	e e:		o o:
Low		a a:	

- **Hausa diphthongs**:

Hausa has various consonants because of the nearness of glottalic arrangement appearing differently in relation to the voiceless and voiced ones and furthermore palatalized and labialized velars close by basic ones. The glottalic consonants are acknowledged as implosives (ɓ, ɗ), ejectives (k’, $k’^w$, $k’^j$, s’) and a glottalized coast (’j). Hausa does not have a "p" sound. It has two particular rhotics: a retroflex flap and an apical tap or roll. All consonants can happen as geminates (double consonants) (Yahaya, 1988). Table 4.2 shows the Hausa language consonants.

Table 4.2: Hausa language consonants



		Labial	Dental-Alveolar	Retroflex	Palatal	Velar	Labio-Velar	Palato-Velar	Glottal
Stop	Voiceless		t			k	k^w	k^j	ʔ
	Voiced	b	d			g	g^w	g^j	
	Glottalic	ɓ	ɗ		‘j	k’	$k’^w$	$k’^j$	
Affricate	Voiceless				tʃ				
	Voiced				dʒ				
Fricative	Voiceless	f f^j	s		ʃ				h
	Voiced		z						
	Glottalic		S’					ʔj	
Nasal		m	n		ŋ				
Liquid			l r	ɽ					
Glide					j		w		

From Table 4.2, a "j" superscript shows palatalization while a superscript "w" demonstrates labialization. The three tones exhibit in Hausa can be classified as high, low and falling tone. The tones are indicated in linguistic work yet are not indicated in standard orthography. In semantic works, low tone is set apart by a grave complement, high tone with an intensifying punctuation (or left unmarked), and falling tone with a circumflex.

- **Script and Orthography**.

Earlier Hausa language was first written in Arabic script (ajami) and latter in 19th century, during the colonization of countries like Nigeria by Britain, Hausa written script is replaced with Latin (boko). Although “ajami” is still used in Quranic education and in some poetry, “boko” has taken over the media and western education sector. In “boko”, the tone and vowel length are not marked, and the difference between the two rhotics is ignored. “Boko” has 30 letters (23 consonants, 5 vowels, and 3 diphthongs) as shown in Table 4.1 and 4.2. Many of the basic sounds are represented by digraphs, which are usually handled as sequences of two letters in the alphabetic order.

- **Lexicon**

Hausa has absorbed a vast number of loanwords, mostly the Arabic words. In the semantic spheres of religion, government, administration, and literature, words of Arabic origin are predominant. More recently, English had a pervasive influence in Nigeria, while the same happened with the French in Niger. Other contributors include the Nigerian and Nigerien languages such as Kanuri, Yoruba and Fulani, and the North African like Mande and Tuareg. Hausa, like many African languages, has a special class of words with particular sound characteristics, called ideophones, associated with vivid sensory or mental experiences.

4.2 Text Corpus Development

Due to the lack of the existing speech database that is suitable for this research, a recorded speech database is built in this research. The development of recorded speech database begins with the text corpus. In this research, the text is extracted from major Hausa daily newspapers, which includes, www.bbc.com/hausa, www.voahausa.com, www.hausa.leadership.ng, www.aminiya.dailytrust.com.ng, as well as reading novels such as Jiki magayi, and Magana jari ce. On top of that, some of the sentences were composed by a linguist based on the day-to-day usage. A total of 401 sentences were collected, containing about 2,350 words. After the manual collection, to ensure phonetically richness and balance, with a good mixture of Hausa words, phones, and syllables, a prompt selection was used using the HMM-based toolkit to select not only the phonetically rich and balance sentences but also ensures the reasonable length for easy recording. At the end of this process, a text corpus containing 179 phonetically rich and balances sentences with about 1,046 words count (including the repetitive words).

4.3 Speech Corpus Development

Development of the speech corpus is one of the difficult tasks in the TTS system development and it requires several stages, which includes, sentence creation, speech recording, pronunciation dictionary, segmentation and labeling, and validation of recorded speech and transcription.

4.3.1 Sentence Creation

Building a rich speech database requires the construction of phonetically rich and balanced sentences from the target language. Phonetically rich and balanced sentences refer to sentence that covers all phonemes of the target language and has adequately been represented in the texts. A text of about 1,046 words was collected from different sources of Hausa language which covers 47.5% from Hausa daily newspapers, 31.3% from Hausa novels, and 21.2% from linguist based on the day-to-day usage. For good distribution of the phonemes, free of grammatical error, and the presence of any non-Hausa word, the sentences were validated by a linguistic expert, from which 179 sentences were selected that has a good mixture of Hausa words, syllables, and phones. The sentences are within the range of five to seven words. The sentences have a total of 1,046 words, 3,104 syllables and 7325 total numbers of phonemes. 29% of the sentences consist of 7 words, 26.3% consists of sentences with 6 words and 44.7% consists of 5 words.

4.3.2 Speech Recording

The speech was conducted in a software engineering research laboratory. In order to develop a good speech database, the recording was carried out by three different Hausa native speakers at different places, but only the speeches from one of them will be selected.

The first speaker is aged 30 years old and the recording was conducted in a software engineering laboratory using the following set of equipment; a speedstar laptop computer (Intel core i7, 2.66GHz processor, 4GB ram, and 500GB HDD), stereo headset (with regular microphone) and Praat software. During the recording process, all the air conditions were turn off to minimize the external noises, the noise of the computer fan was also taken into consideration by asking the speaker to set adequate distance for the noise sources. The microphone was kept at a distance of about 10cm to avoid lips smack, the speaker was given ample time to reads and practice the sentences beforehand, so to get him acquainted with the words in the sentences. The recording process took about three hours and the total time of recorded speech was 21 minutes 50 seconds, the sampling rate of recording is 16 kHz and recorded speech was saved as ".wav" file format with 16 bit. This recording rate has been widely used for speech recording analysis (Stolcke, Mandal, & Shriberg, 2012). One of the unique characteristics of Praat software used for the recording is that it provides different types of speech analysis features and very easy to handle. The second speaker is 33 years old, and the recording was conducted at 2effect studio, Kano, Nigeria using rode compact microphone and HP laptop computer, with the microphone was set at a distance 20cm. The recording settings are the same as the first recording. The duration of the second recording is 25 minutes 30 second. The third speech recording was conducted in Zongolink Hausa radio studio London using Asio echo firewire microphone with the same recording settings as the previous recordings.

The three recordings were evaluated, and only one speaker's speech is selected for the development of the Hausa TTS system. The speech data that was recorded in 2effect studio at Kano, Nigeria (the second speaker) was selected based on the listening evaluation conducted with native Hausa listener. This speech data was selected because it has minimal background noise, clearer sound, and proper pronunciation of sentence, words or syllable. Some of the difficulties encountered during this recording were the difficulty in obtaining professional speaker and the time needed for the recording especially the recording performed in Kano, Nigeria.

Recorded speech data and its transcription (segmented speech data) require the validation to ensure a rich speech corpus that is free from poor utterances and has a proper phonemes distribution. It is used to trim the bad utterances and to authenticate the distribution of the phonemes in the sentences and transcription, so as to ensure excellent quality of phoneme distribution. It is performed towards the end of resource accumulation process. At the end of resource accumulation process, a professional Hausa native speaker not involved during the recording process and transcription was involved to evaluate the recorded speech data and its transcription.

Figure 4.1 shows the snapshot of recording session conducted in Software Engineering Laboratory, while Figure 4.2 depicts the snapshot of recording session conducted in

2effect studio Kano, and Figure 4.3 shows the snapshot of recording session conducted at ZongoLink Hausa radio, London.

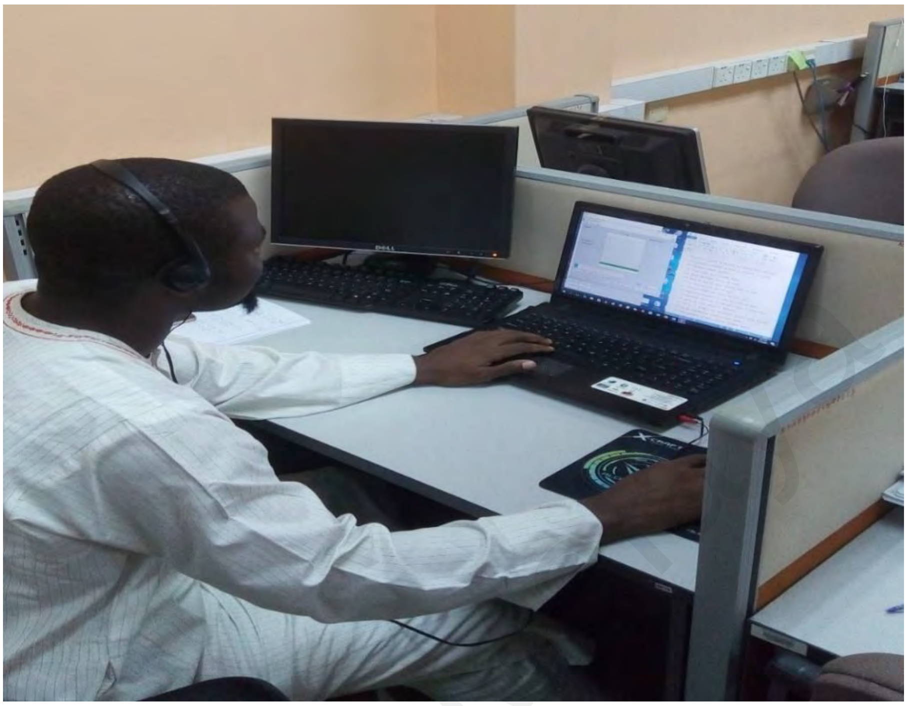


Figure 4.1: A Snapshot of recording session conducted in Software Engineering Laboratory

Figure 4.2: A Snapshot of recording session conducted in 2effect studio Kano

Figure 4.3: A Snapshot of recording session conducted at ZongoLink Hausa radio, London

Table 4.3: Summary of Hausa speech recording

Speaker	No of sentences	Size of Recorded Speech	Remarks
Speaker 1	179 sentences	21 min., 50 sec.	The quality of recorded speech was bad due to the presence of background noises as the recording was conducted in a software engineering laboratory. (Rejected)
Speaker 2	179 sentences	25 min., 30 sec.	It has very minimal background noise with very high quality, clear sound and proper pronunciation. (Selected)
Speaker 3	179 sentences	23 min., 31 sec.	It also has a very minimal background noise, good quality but the pronunciation of some words are not clear. (Rejected)

4.3.3 Pronunciation Dictionary

Pronunciation dictionary is one of the fundamental processes that defines the combination of phoneme during the synthesis process. The pronunciation dictionary is usually prepared manually, which is difficult to create, especially for a large vocabulary, and a variety of possible pronunciation (Biadsy, Habash, & Hirschberg, 2009). Pronunciation dictionary is also used to transcribe a particular language to represent the combination of phonemes that made up a word. In a pronunciation dictionary, the same word may contain different pronunciation information, depending on the context it was created for. The pronunciation dictionary in this research was manually created by a linguistic and was evaluated by another linguistic to ensure that the pronunciation is error free. A total of 542 words selected from 179 sentences used for the creation of text and speech corpus was transcribed. Table 4.4 depicts some sample of the pronunciation dictionary for Hausa.

Table 4.4: Sample list of pronunciation dictionary for Hausa

Word	Pronunciation	Meaning in English
ABINCI	(aa) (b iy n) (c ih)	Food
ADDINI	(ax d) (d iy) (n iy)	Religion
BIYAR	(b ih) (y aa r)	Five
GABAS	(g aa) (b ax s)	East
RABI	(r aa) (b iy)	Half

4.3.4 Segmentation and Labelling

Segmentation and labelling is one of the tedious tasks, not only for under-resourced languages but also for well-resourced language. A recorded speech can be manually or automatically segmented. Manual segmentation and labelling of recorded speech is not only time consuming and expensive but also slow down the development processes for TTS system based on the unit selection and HMM-based synthesis. Manual segmentation of the recorded speech of one speaker may not be applicable to another speaker (Kim & Conkie, 2002). Automatic segmentation provided by some segmentation tool such as

HTK using Viterbi alignment algorithm has shown a promising result for non-tonal languages, which also provides reliable phonetic alignment (Mporas, Lazaridis, Ganchev, & Fakotakis, 2009). These tools, however, require an initial speech acoustic model of a particular language to perform the automatic alignment (Besacier, Barnard, Karpov, & Schultz, 2014). Many of the under-resourced languages including Hausa do not have the existing speech acoustic model for performing the automatic segmentation. Due to the absence of existing Hausa speech corpus that can be used for automatic segmentation, a total of 50 sentences was manually segmented and labeled by a linguistic expert. To ensure proper segmentation, a three-stage segmentation was conducted: (1) segmentation based on words, (2) segmentation based on syllabus and (3) segmentation based on phoneme level. Figure 4.4 depicts the sample of segmentation and labeling using Praat software and Table 4.5 shows the summary of resource accumulation.

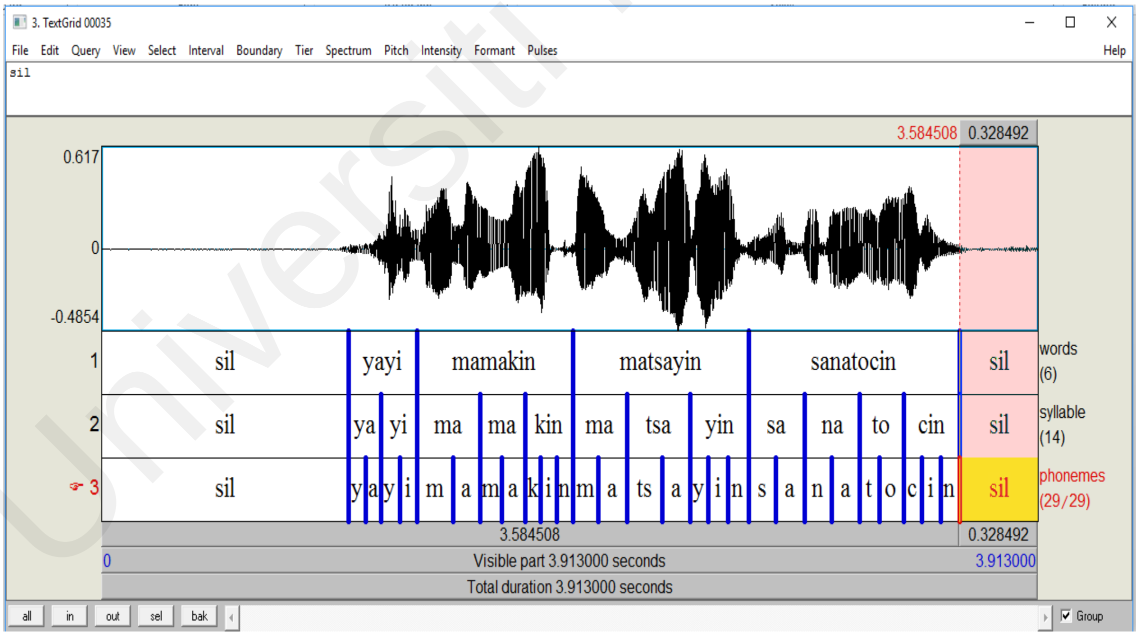

3. TextGrid 00035
File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses Help
sil
3.584508 0.328492
0.617
0
-0.4854
1 sil yayi mamakin matsayin sanatocin sil words (6)
2 sil ya yi ma ma kin ma tsa yin sa na to cin sil syllable (14)
3 sil y a y i m a m a k i n m a ts a y i n s a n a t o c i n sil phonemes (29/29)
3.584508 0.328492
0 Visible part 3.913000 seconds 3.913000
Total duration 3.913000 seconds
all in out sel bak
☑ Group

Figure 4.4: Sample segmentation and labelling using Praat software "Yayi mamakin matsayin sanatocin/He was surprise with senate decision"

Table 4.5: Summary of resource accumulation

Required Resources	**Purpose**	**Strategies**	**Size of data**	**Sources/Conducted**	**Validation**
Text collection	For creation of text corpus	Manual collection from online Hausa newspaper, novels and short stories	401 sentences	bbc.com/hausa voahausa.com hausa.leadership.ng Jiki magayi Magana jari ce /Abubakar Ahmad	Linguist
Text Corpus	Phonetically rich and balanced sentences for speech corpus development	Automatic process using prompt selection	179 sentences	Extracted text from daily news and other	Linguist
Speech Corpus	Collection of recorded speech	Recording by a male native speaker in a noise free environment	179 sentences	Hausa native speaker	Linguist
Transcriptions	The transcription of speech providing the linguistic information	Manually transcribed using Praat software	50 sentences	Linguist	Linguist
Pronunciation Dictionary	Enable the system to synthesize words, syllabus or phrase not present in the database	Grapheme-to-phoneme rules	words	Linguist	Linguist

4.4 Chapter Summary

In this chapter, the process of building the Hausa speech database was discussed. The resource accumulation includes Hausa text corpus, Hausa speech corpus, pronunciation dictionary, and transcription. This chapter also explained the processes involved during the recording, segmentation, and labeling. A total of 179 sentences consisting of 1,046 words were recorded and 50 sentences were segmented and label manually for the training data.

CHAPTER 5: DEVELOPMENT OF HAUSA TTS SYSTEM

This chapter discusses the development process of the Hausa TTS system.

5.1 Development of Hausa TTS System

Acoustic models provide the statistical representations of distinct sounds that make up a word. The development of Hausa acoustic model was performed using the proposed cross-lingual adaptation and DNN-based synthesis. The cross-lingual adaptation was initially applied to HMMs for under-resourced languages, which yields a better performance than the other techniques used for the development of TTS system for the under-resourced languages. With the improved performance of DNN-based speech synthesis system, especially in terms of naturalness and intelligibility, this research experiments the cross-lingual adaptation for DNN-based synthesis. The development of the acoustic model consists of several stages, such as features extraction, time-alignment, normalization, forced-alignment, and training DNN. Figure 5.1 depicts the framework for the development of speech acoustic model of Hausa TTS system.

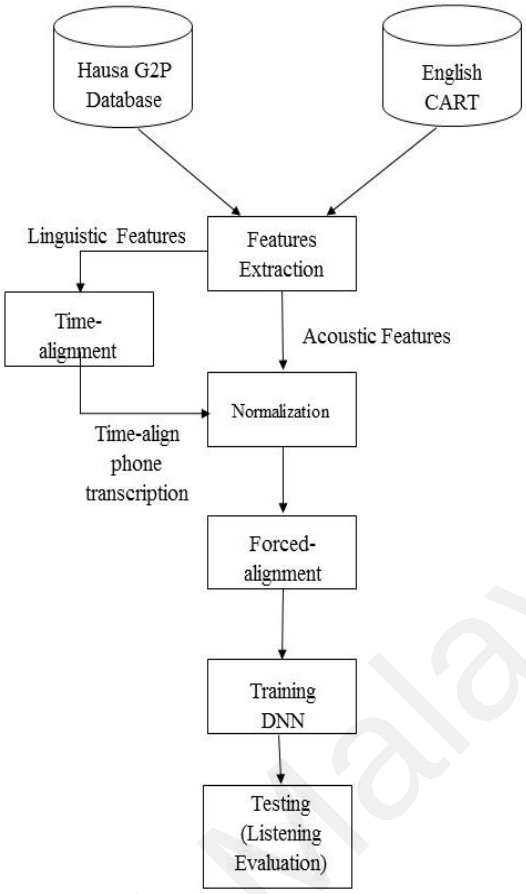


Figure 5.1: Development framework of Speech Acoustic Model for Hausa TTS System

5.1.1 TTS Development Environment

Some of the toolkits used for the development of Hausa TTS system include Matlab, Praat, Straight vocoder etc.

- MATLAB : MATLAB is a language of technical computing that is used for computational intensive tasks, where it has components that support signal processing, data analysis and visualization, and so on
- Praat – It is free software for sound manipulation, phonetic and acoustic analysis.
- STRAIGHT – It stands for Speech Transformation and Representation using Adaptive Interpolation weiGHTed spectrum, which is used for the conversion of wave signals into speech.

- SAPI – SAPI stands for Speech Application Programming Interface. It is an application programming interface that allows the use of speech synthesis and speech recognition within the windows applications.

5.1.2 Features Extraction

In statistical parametric speech synthesis, feature extraction of the training data is the most important part of the training process. The linguistic feature for the DNN synthesis consists of the questions about the linguistic contexts of phoneme and its numerical values (e.g. relative position of the current frame in the current phoneme, duration of the phoneme, and so on) (Zen et al., 2013). The acoustic output feature is the MFCC and log F0. The input linguistic feature for Hausa was generated from the proposed cross-lingual approach (i.e time-alignment). The output acoustic feature is 40 mel-cepstral coefficient, fundamental frequency, and filterBank energy (extracted every 5 milliseconds from the recorded speech database sampled at 16 kHz).

5.1.3 Time-aligned Phone Transcription

One of the advantages of the neural network is its ability to learn and use the existing linguistic features to predict the output of the acoustic features at any given frame in time (Gutkin, Ha, Jansche, Kjartansson, et al., 2016). The data in this research consists of the phoneme level transcriptions without time-alignment and the speech data for each phoneme transcribed in plain orthography at the utterance level (Watts, Ronanki, Wu, Raitio, & Suni, 2015). The input text were analyzed and normalized to generate the phonetic and contextual representation using a slightly modified text processing module of Kaldi for the purpose of the DNN training (Potard, Aylett, Baude, & Motlicek, 2016). The normalized data were then time-aligned with the phone duration of English festival using the dynamic time warping algorithms to generate the time-aligned phone transcription.

During the training, at the input layer, a mapping between the Hausa and English phonetic transcription was performed using the dynamic programming to find the optimal global alignment between both Hausa and English strings. The phoneme /fy/ and /ts/ was added to the exceptional list as these two are absent in the English database. At the hidden layer, a transfer learning was performed using the learning hidden unit contributions approach. Figure 5.2 shows the framework for Hausa TTS system using cross-lingual approach and DNN-based synthesizer.

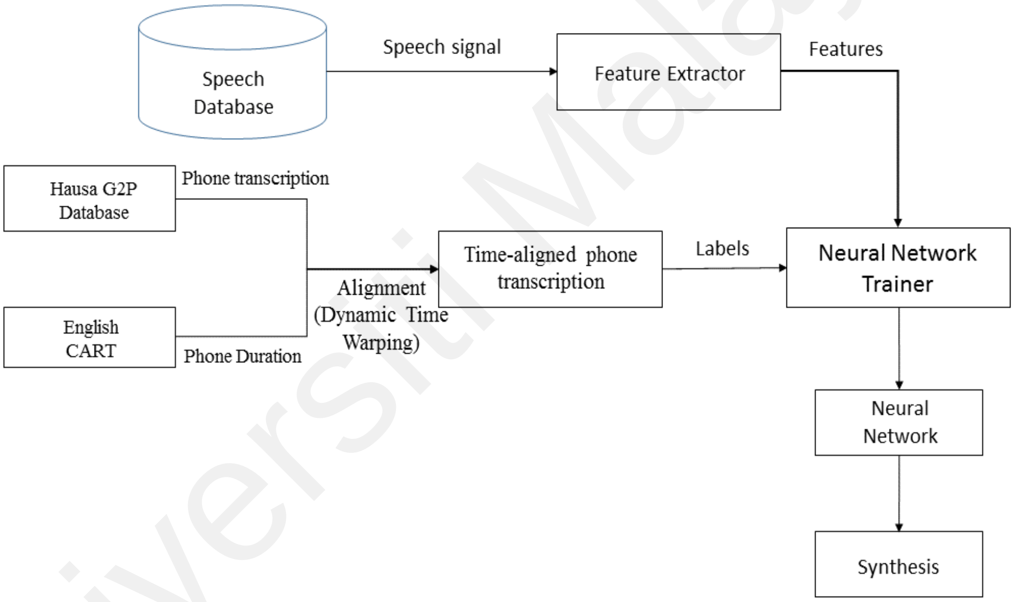


Figure 5.2: Framework for Hausa TTS system using cross-lingual and DNN techniques.

- **Mapping the Hausa and English phonemes**

Phoneme mapping can be used in cross-lingual adaptation to reduce the word error rate and improve the accuracy of the TTS system. In this research, the English language was selected as the source language because of its phoneme similarities with The Hausa language. A phoneme mapping proposed by Wu et al. (2008) was adopted in this research to map the Hausa phonemes to the nearest English phoneme.

Table 5.1 illustrates the mapping between the Hausa and English phonemes, in which the Hausa phonemes that are not available in English language and extra English phonemes are disregarded.

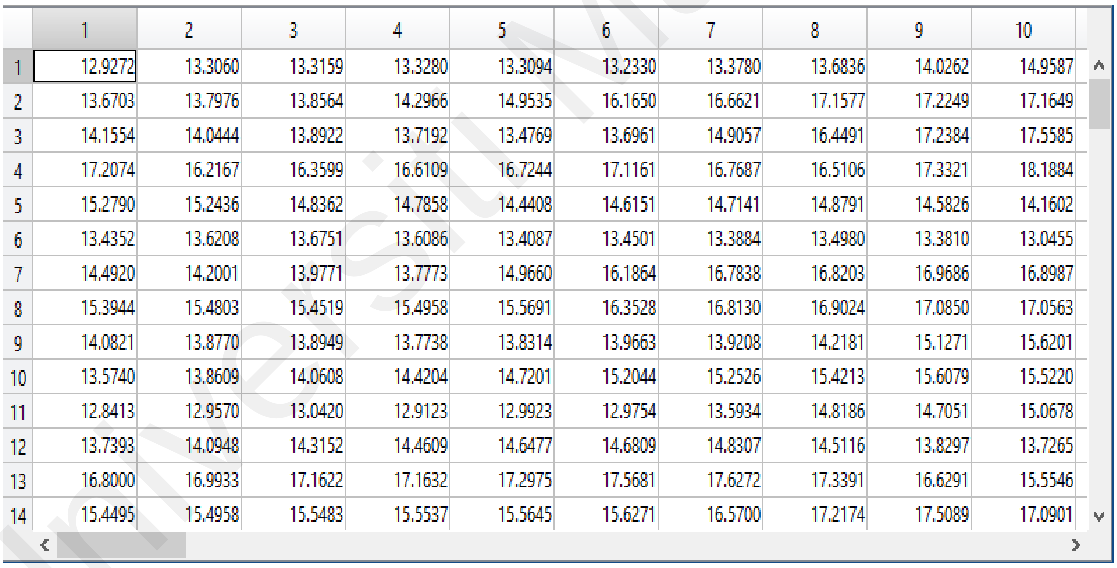
Table 5.1: Hausa and English phoneme mapping

Phoneme Type	English	English	Hausa	Hausa
Consonant	/b/	**b**at	/b/	**b**iyu
	/d/	car**d**	/d/	au**d**u
	/ð/	brea**the**	/d/	ka**d**a
	/dʒ/	ba**dge**	/j/	**je**ji
	/f/	**f**an	/f/	**f**ata
	/g/	ba**g**	/g/	**g**ida
	/h/	**h**uddle	/h/	she**h**u
	/j/	**y**am	/j/	**j**ayayya
	/k/	**c**ar	/k/	**k**are
	/l/	**l**eg	/l/	**l**iman
	/m/	**m**ile	/m/	**m**akaranta
	/n/	**n**ight	/n/	su**n**a
	/r/	w**r**ite	/r/	**r**ashawa
	/s/	**s**ight	/s/	**s**arki
	/ʃ/	**s**hy	/sh/	**sh**awara
	/t/	**t**eeth	/t/	ci**t**ura
	/tʃ/	**ch**ina	/c/	**c**ewa
	/w/	**w**ine	/w/	ce**w**a
	/z/	**z**ibra	/z/	ar**z**iki
	/ʒ/	vi**s**ion	/sh/	**sh**ugaba
	/θ/	**th**ings	/t/	fe**t**ur
Vowel	/a:/	f**a**ther	/aa/	b**a**shi
	/a/	f**a**t	/a/	s**a**rki
	/ɪ/	b**i**t	/ih/	b**i**yar
	/e/	b**e**ll	/e/	b**e**llo
	/e:/	f**u**r	/u/	s**u**na
	/uː/	s**oo**n	/uw/	ab**u**bakar
	/ʊ/	Sh**oe**	/uw/	w**u**ta
	/Ɔ:/	b**oo**k	/Ɔ:/	aud**u**
	/iː/	feel	/ih/	sark**i**
	/ɒ/	b**o**ne	/ow/	b**o**ko
Dip-thongs	/ai/	r**i**de,	/ai/	**ai**ki
	/au/	d**ow**n	/aw/	**au**du
	/ui/	**coi**n	/ui/	gw**ui**wa

5.1.4 Normalization

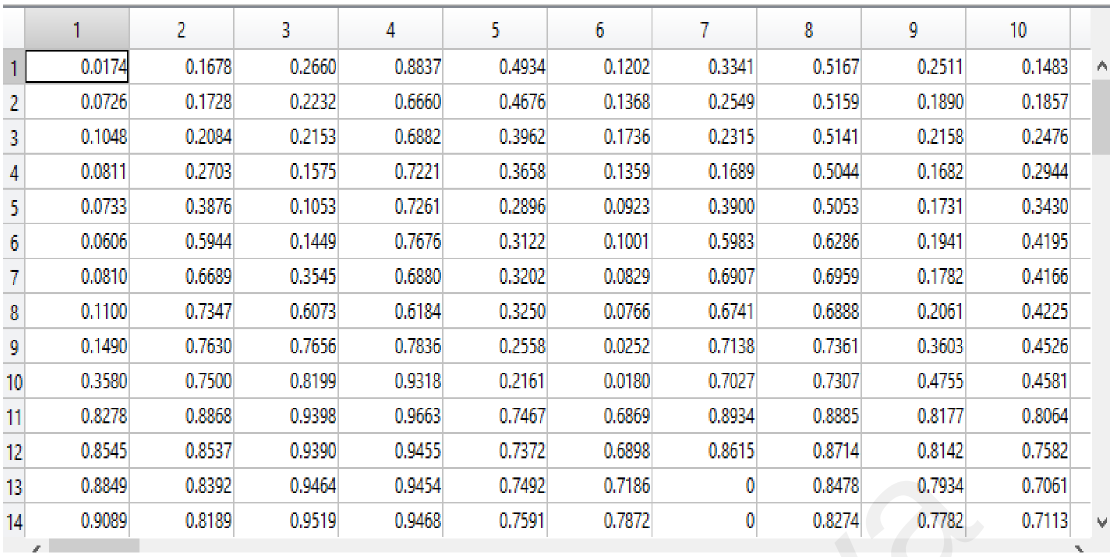
Feature scaling is a technique used to standardize the range of the independent variables or the features of the data. In data processing, it is also known as data normalization and is generally performed during the data pre-processing stage. Training the deep neural network requires the normalization of the extracted features to increase the speed of the training, and reducing the training error. Generally, before training, any neural network must undergo the normalization process. This is because neural network uses the binary features and these features were all set to the range of 0 and 1. The process of normalization is categorized into two; the min-max and mean-variation (Wu, Watts, & King, 2016). The min-max normalize features to the range of [0.01 – 0.99], while the mean-variances normalize features of zero mean and unit variance. In this research, the output acoustic features were normalized using min-max to the range of [0.01 – 0.99] and the input linguistic features were normalized to zero mean and unit variance.

Figure 5.3 depicts the sample of MFCC feature extracted before normalization while figure 5.4 depicts the features after the normalization.



	1	2	3	4	5	6	7	8	9	10
1	12.9272	13.3060	13.3159	13.3280	13.3094	13.2330	13.3780	13.6836	14.0262	14.9587
2	13.6703	13.7976	13.8564	14.2966	14.9535	16.1650	16.6621	17.1577	17.2249	17.1649
3	14.1554	14.0444	13.8922	13.7192	13.4769	13.6961	14.9057	16.4491	17.2384	17.5585
4	17.2074	16.2167	16.3599	16.6109	16.7244	17.1161	16.7687	16.5106	17.3321	18.1884
5	15.2790	15.2436	14.8362	14.7858	14.4408	14.6151	14.7141	14.8791	14.5826	14.1602
6	13.4352	13.6208	13.6751	13.6086	13.4087	13.4501	13.3884	13.4980	13.3810	13.0455
7	14.4920	14.2001	13.9771	13.7773	14.9660	16.1864	16.7838	16.8203	16.9686	16.8987
8	15.3944	15.4803	15.4519	15.4958	15.5691	16.3528	16.8130	16.9024	17.0850	17.0563
9	14.0821	13.8770	13.8949	13.7738	13.8314	13.9663	13.9208	14.2181	15.1271	15.6201
10	13.5740	13.8609	14.0608	14.4204	14.7201	15.2044	15.2526	15.4213	15.6079	15.5220
11	12.8413	12.9570	13.0420	12.9123	12.9923	12.9754	13.5934	14.8186	14.7051	15.0678
12	13.7393	14.0948	14.3152	14.4609	14.6477	14.6809	14.8307	14.5116	13.8297	13.7265
13	16.8000	16.9933	17.1622	17.1632	17.2975	17.5681	17.6272	17.3391	16.6291	15.5546
14	15.4495	15.4958	15.5483	15.5537	15.5645	15.6271	16.5700	17.2174	17.5089	17.0901

Figure 5.3: Sample of MFCC before normalization



	1	2	3	4	5	6	7	8	9	10
1	0.0174	0.1678	0.2660	0.8837	0.4934	0.1202	0.3341	0.5167	0.2511	0.1483
2	0.0726	0.1728	0.2232	0.6660	0.4676	0.1368	0.2549	0.5159	0.1890	0.1857
3	0.1048	0.2084	0.2153	0.6882	0.3962	0.1736	0.2315	0.5141	0.2158	0.2476
4	0.0811	0.2703	0.1575	0.7221	0.3658	0.1359	0.1689	0.5044	0.1682	0.2944
5	0.0733	0.3876	0.1053	0.7261	0.2896	0.0923	0.3900	0.5053	0.1731	0.3430
6	0.0606	0.5944	0.1449	0.7676	0.3122	0.1001	0.5983	0.6286	0.1941	0.4195
7	0.0810	0.6689	0.3545	0.6880	0.3202	0.0829	0.6907	0.6959	0.1782	0.4166
8	0.1100	0.7347	0.6073	0.6184	0.3250	0.0766	0.6741	0.6888	0.2061	0.4225
9	0.1490	0.7630	0.7656	0.7836	0.2558	0.0252	0.7138	0.7361	0.3603	0.4526
10	0.3580	0.7500	0.8199	0.9318	0.2161	0.0180	0.7027	0.7307	0.4755	0.4581
11	0.8278	0.8868	0.9398	0.9663	0.7467	0.6869	0.8934	0.8885	0.8177	0.8064
12	0.8545	0.8537	0.9390	0.9455	0.7372	0.6898	0.8615	0.8714	0.8142	0.7582
13	0.8849	0.8392	0.9464	0.9454	0.7492	0.7186	0	0.8478	0.7934	0.7061
14	0.9089	0.8189	0.9519	0.9468	0.7591	0.7872	0	0.8274	0.7782	0.7113

Figure 5.4: Sample of MFCC after normalization

5.1.5 Forced-alignment

Forced-alignment is the most commonly used techniques for the training of an under-resourced language (Mullah et al, 2015; Maas et al., 2017). Forced-alignment based on HMM-DNN (Maas et al., 2017) was applied in this research. The training data used in this research, consists of phonetic and contextual representation, phone duration, MFCC, FBE and log F0. A senone label is assigned to each acoustic output frame in each training utterance. A forced alignment of the ground-truth transcriptions is used to generate a sequence of senone labels for each utterance, which is consistent with the word transcription for the utterance. The aligned data is then used to train a neural network acoustic model. In this research, a single forced alignment was adopted to produce the baseline system, which is similar to the work done by Mullah et al. (2015).

5.1.6 Training the Speech Acoustic Model based on DNN

The speech data used for the DNN training consists of 50 utterances. The input features include the binary features for categorical linguistic contexts and their numerical features, while the output features consist of log F0 value, 40 mel-cepstral coefficient, and band 5 aperiodicity. The training was performed using a feed forward back propagation neural network, which consists of three hidden layers each consists 50, 25, 25 units respectively, and a network consist of two hidden layers with 50 and 25 units. Figure 5.5 and 5.6 depicts the training performance of the data used for 3 hidden layers and 2 hidden layers.

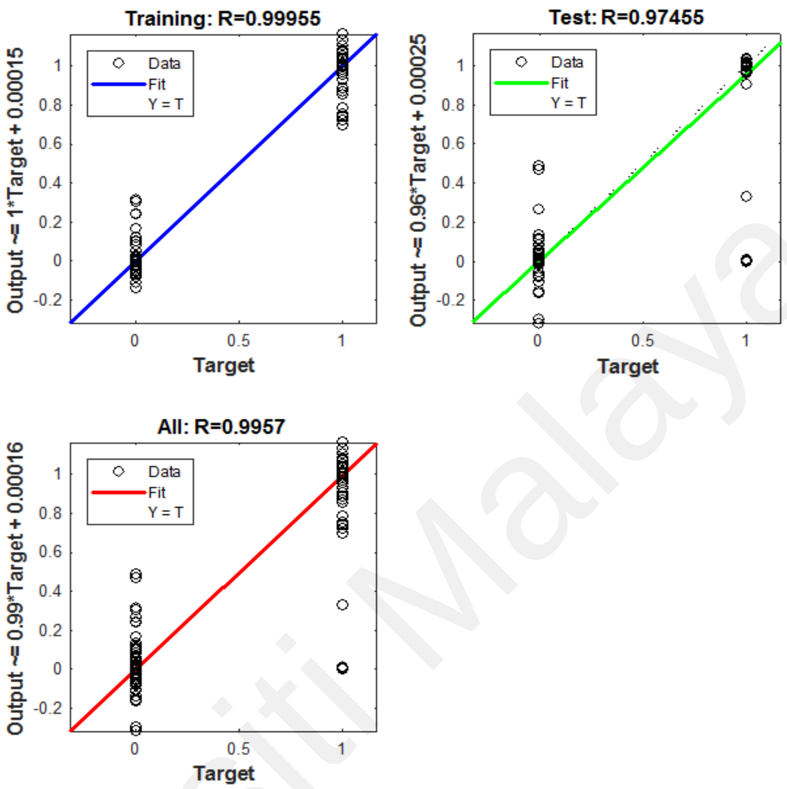


Figure 5.5: 3 hidden layers performance at 233 epochs

From figure 5.5, it shows that the training performance was 99.96% approximately, the testing performance was 97.46% and overall performance for the training was 99.57% for the three hidden layer which is better than the two hidden layers.

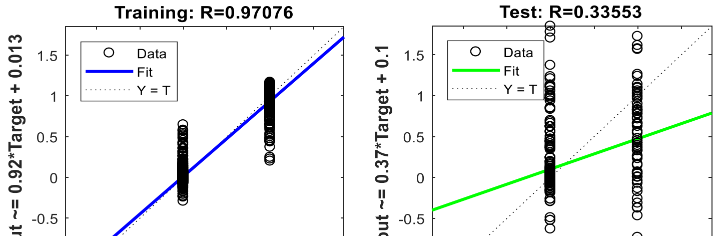


Figure 5.6: 2 hidden layers performance at 291 epochs

From figure 5.6, it shows that the training performance was 97.10% approximately, the testing performance was 33.55% and overall performance for the training was 81.81% for the two hidden layer.

Table 5.2: Summary of the training result

No. of Layers	No. of Neurons	Epochs	Accuracy	Misclassification
3 Hidden	100	233	99.57%	0.43%
2 Hidden	75	291	81.81%	18.19%

The overall performance of the training using three hidden layers is 99.57% and is obtained at 233 epochs while the overall performance using 2 hidden layers is 81.81% obtained at 291 epochs. This shows that the data training using three hidden layers outperformed the training using two hidden layers. The 3 layer is applied for the Hausa TTS system due to its highest accuracy. Figure 5.7 and 5.8 depict the development environment for the training and synthesis. The part of source codes for the pre-processing, training DNN and synthesis stages are shown in appendix E

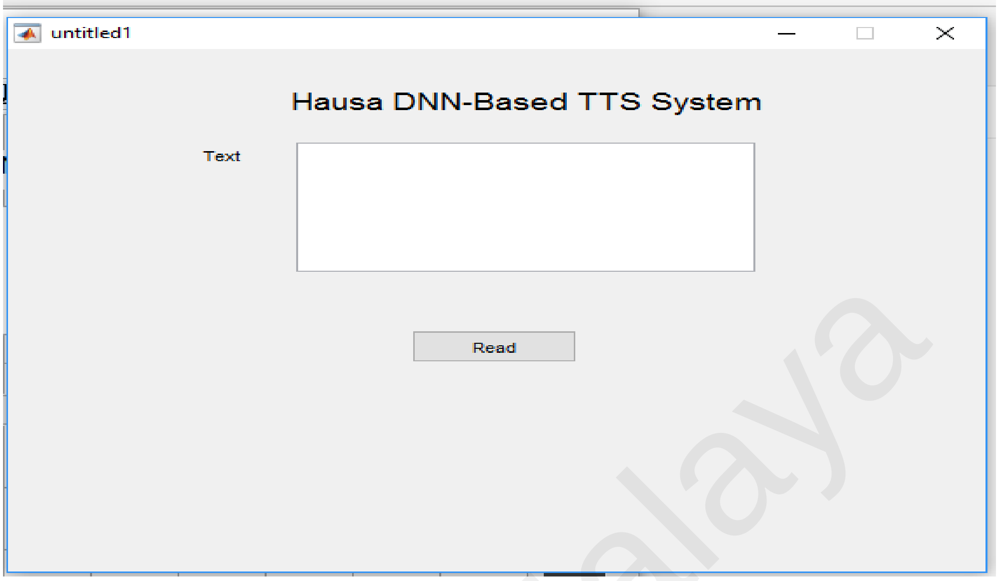
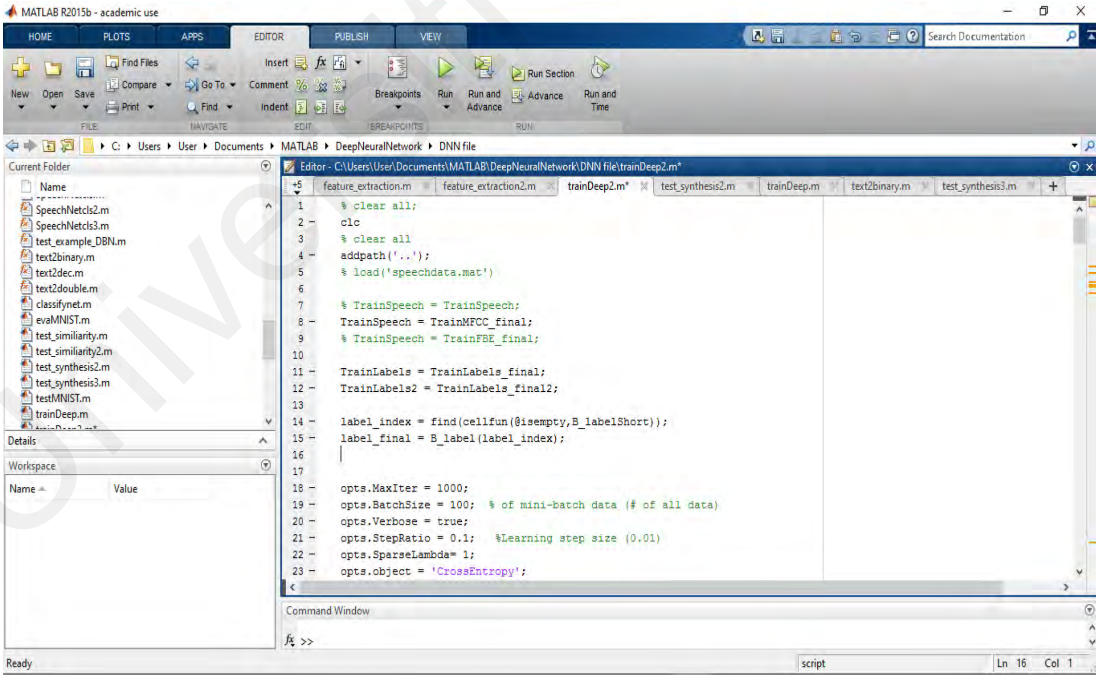


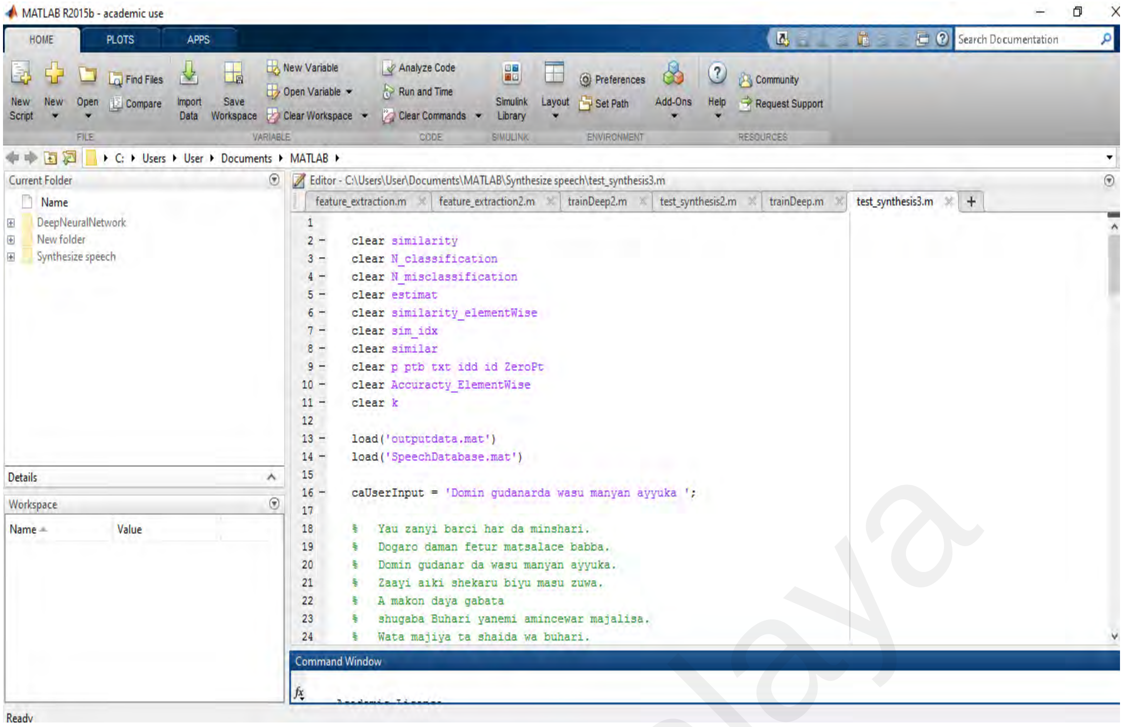
Figure 5.7: Front-end of Hausa TTS system



```
% clear all;
clc
% clear all
addpath('..');
% load('speechdata.mat')

% TrainSpeech = TrainSpeech;
TrainSpeech = TrainMFCC_final;
% TrainSpeech = TrainFBE_final;

TrainLabels = TrainLabels_final;
TrainLabels2 = TrainLabels_final2;

label_index = find(cellfun(@isempty,B_labelShort));
label_final = B_label(label_index);


opts.MaxIter = 1000;
opts.BatchSize = 100;  % of mini-batch data (# of all data)
opts.Verbose = true;
opts.StepRatio = 0.1;   %Learning step size (0.01)
opts.SparseLambda= 1;
opts.object = 'CrossEntropy';
```

```
>>
```

Figure 5.8: Development environment for the training



```

clear similarity
clear N_classification
clear N_misclassification
clear estimat
clear similarity_elementWise
clear sim_idx
clear similar
clear p ptb txt idd id ZeroPt
clear Accuracty_ElementWise
clear k

load('outputdata.mat')
load('SpeechDatabase.mat')

caUserInput = 'Domin gudanarda wasu manyan ayyuka ';

%   Yau zanyi barci har da minshari.
%   Dogaro daman fetur matsalace babba.
%   Domin gudanar da wasu manyan ayyuka.
%   Zaayi aıki shekaru biyu masu zuwa.
%   A makon daya gabata
%   shugaba Buhari yanemi amincewar majalisa.
%   Wata majiya ta shaida wa buhari.
```

Figure 5.9: Development environment for the synthesis

5.2 Chapter Summary

In this chapter, the development of the proposed DNN-based cross-lingual adaptation was discussed and it begins with the feature extraction. This chapter also explains the normalization, time-alignment, forced alignment, and DNN training stages. The product of the cross-lingual adaptation and DNN training show a good performance in terms of the training result with about 99.57% accuracy using the three hidden layers.

CHAPTER 6: EVALUATION, RESULTS AND DISCUSSIONS

This chapter discusses the evaluation of the developed Hausa TTS system. This chapter also explains the criteria used for the selection of the test sentences and evaluation subjects. The result of the evaluation was also discussed.

6.1 Respondents for Listening Evaluation

The developed Hausa TTS system was evaluated for its intelligibility and naturalness using the 50 sentences that were not part of the training. The listening evaluation involved 50 Hausa native listeners that consist of 10 respondents from University Malaya, 20 respondents from the International Islamic University, and another 20 respondents from Universiti Putra Malaysia as shown in table 6.1. The evaluators were fluent Hausa speakers.

Some of the criteria used for the selection of respondents for listening evaluation are as follows;

- Respondents must be a native Hausa speaker.
- Respondents can be a male or female within the range of 18 to 40 years due to their good hearing representation.
- People from broad spectrum; undergraduate or post-graduate (targeting UM, IIUM and UPM).

Table 6.1: Subjects involved in the listening evaluation

Institution	No. of Respondents
UM	10
IIUM	20
UPM	20

6.2 Criteria for Selecting the Sentences for Testing

The criteria used for the selection of the sentences in the evaluation is as follow.

- Sentence that are not part of the sentences used for the training.
- A total of 50 meaningful sentences were selected from the general domain (see appendix C).
- For easy comprehension, only sentences that consist of 4 to 7 words was selected.

6.3 Evaluation Procedures

During the evaluation process, all respondents were given the Hausa synthesized speech to be evaluated on the intelligibility and naturalness. The respondents were given the ample time to listen to the synthesized speech as many times as they want. During the listening evaluation, in order to minimize the external noises, the evaluators were furnished with headphones (a stereo headset with regular microphone) to listen to the synthesized speech in a very conducive and sound proof environments (laboratories, rooms etc).

A special form was designed to aid the evaluation process. The respondents provide their answers and scores on the form given during the evaluation process. All respondents were required to fill-up the evaluation form that includes some personal information (bio data). Respondents were briefed on the purpose of the evaluation as well as the working mechanism of the speech synthesis system and how it works. Respondents were then asked to carefully listen to the synthesized speech (from the laptop used for research) and write what they heard. Respondents were not allowed to know the input text to ensure a fair evaluation of the system. Each respondent listened to five sentences randomly selected from the 50 sentences, write it down, and then answer some questions related to the naturalness and intelligibility.

At the end of the evaluation, a comparison analysis is performed on the performance (i.e. intelligibility and naturalness) of the developed system with the existing system for similar under-resourced development.

6.3.1 Procedures for Intelligibility Evaluation

Intelligibility is the measurement of the ability of the subject evaluators to comprehend the synthesized speech. For intelligibility test, listeners were asked to listen to the synthesized speech and write what they have heard on the form provided. At the end of the evaluation, the response from the respondents is compare with actual sentence to determine the accuracy of the written answer. The speech intelligibility is based on word error rate that includes words deleted, inserted and/or substituted. The evaluators were also asked to rate how difficult it was to comprehend the synthesized speech.

6.3.2 Procedures for Naturalness Evaluation

Naturalness is the measure of the similarity between the human generated speech and the synthesized speech. For evaluating the naturalness, listeners were asked to listen to the synthesized speech and rate how similar the synthetic speech to the human speech using a measuring scale of 5 points Likert scale. The listening procedures are the same with the intelligibility test where the evaluators listen to synthesized speech and select the appropriate score from the 5 points Likert scale.

6.4 Results and Discussion

From the listening evaluation from the 50 respondents, the Hausa TTS system developed using the cross-lingual technique and DNN-based synthesis scores 84% in terms of naturalness and 82% in terms of intelligibility, which is 4.20 and 4.10 of point Likert scale respectively. The proposed technique has also reduced the word error rate (WER) by about 2% better than the existing cross-lingual technique using HMM-based synthesis (at 20 % WER). During the listening test, it was found that the most of the listeners can fully comprehend the synthesized speech after listening to the synthetic speech for two or three times. The listeners indicated that they are satisfied with the synthesized speech, which indicates the improved performance of the DNN-based synthesis than the previous TTS system for the under-resourced language. Figure 6.1 depicts the result of the listening evaluation of the Hausa TTS system. From figure 6.1, it was found that 80% of the respondents rated the synthetic speech to be 80% natural (i.e. 4 of 5 rating) and 20% of the respondents give a 5 upon 5 score (100%) for the naturalness. It also found that the overall performance of the TTS system was acceptable at about 83%. The results of the word error rate show that most of the words that are wrongly comprehended have a mixture of glottalic consonants (such as ɓ, ɗ and so on).

It was also observed that as the length of the sentence become longer, the intelligibility score is negatively affected, as longer sentences contain more words and remembering each of them is increasingly difficult. For the intelligibility test, some of the words that are commonly misspelled include; “tsinci, karamin, wakilai, kauye, and gwamnati” which might be as a result of the absence of some of the phonemes from the resourced language (English).

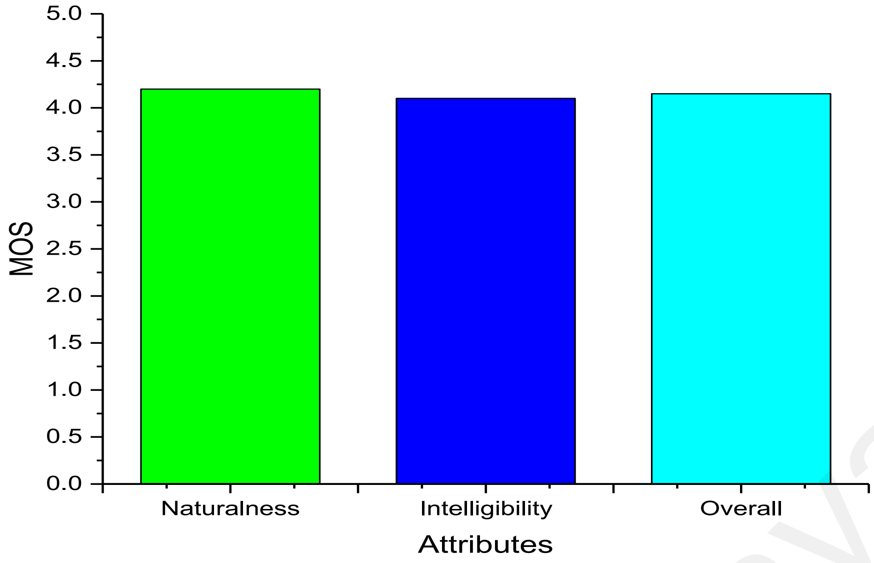


Figure 6.1: Hausa TTS system performance evaluation

Figure 6.2 and 6.3 show the comparative performance in term of intelligibility and naturalness of the DNN-based Hausa TTS system with HMM-based TTS system for some of the under-resourced languages (Sher, et al., 2010; Boothalingam et al., 2013; Mullah et al., 2015). The overall performance of the intelligibility and naturalness of the DNN-based Hausa TTS system was better than the previous HMM-based TTS system for some of the under-resourced languages (Taiwanese, Tamil and Indian Eng. Lan.). It was found that the DNN-based Hausa TTS system was 10% better (the score is 0.2 higher) for its naturalness as compared to some of the HMM-based TTS system developed for other under-resourced languages. In term of intelligibility, the DNN-based Hausa TTS system edge the performance of the existing HMM-based TTS system for other under-resourced languages by 5% (the score is 0.1 higher).

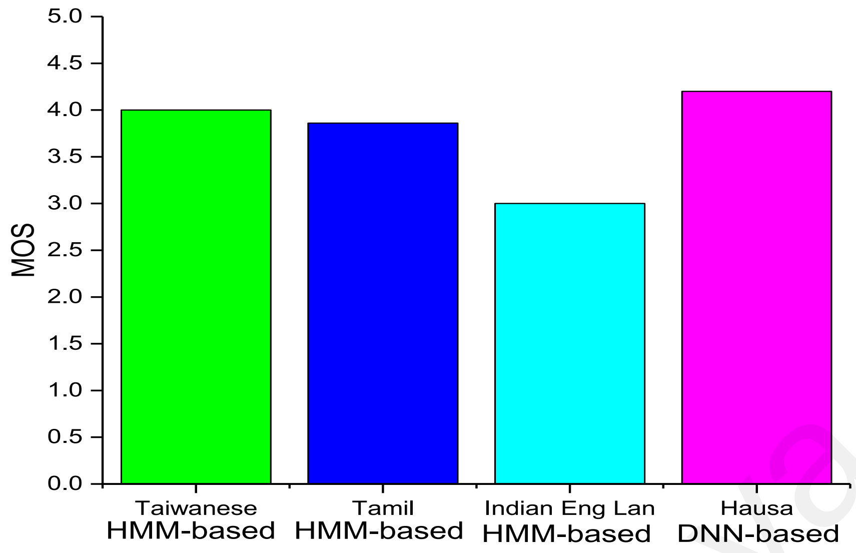


Figure 6.2: Naturalness comparative score between HMM-based and DNN-based TTS systems developed for under-resourced languages

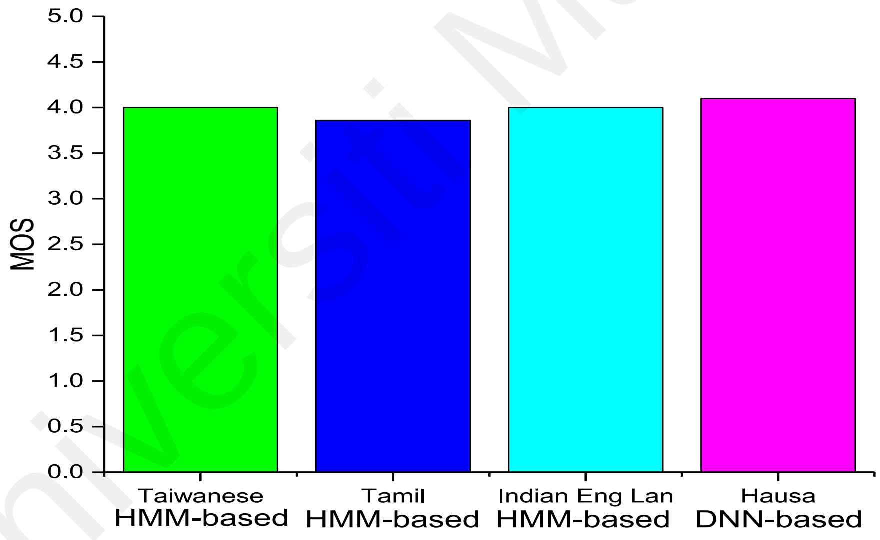


Figure 6.3: Intelligibility comparative score between HMM-based and DNN-based TTS systems developed for under-resourced languages

From figure 6.4, it was found that 20% of the respondents from UM rated the system to be 5 out of 5 and 80% rated it to be 4 out of 5, 15% of the respondents from IIUM rated the system to be 5 out of 5 and 85% rated it to be 4 out of 5 while 25% of the respondents from UPM rated the system to be 5 out of 5 and 75% rated it to be 4 out of 5 for naturalness respectively. Figure 6.4 depicts the comparative naturalness score based on universities.

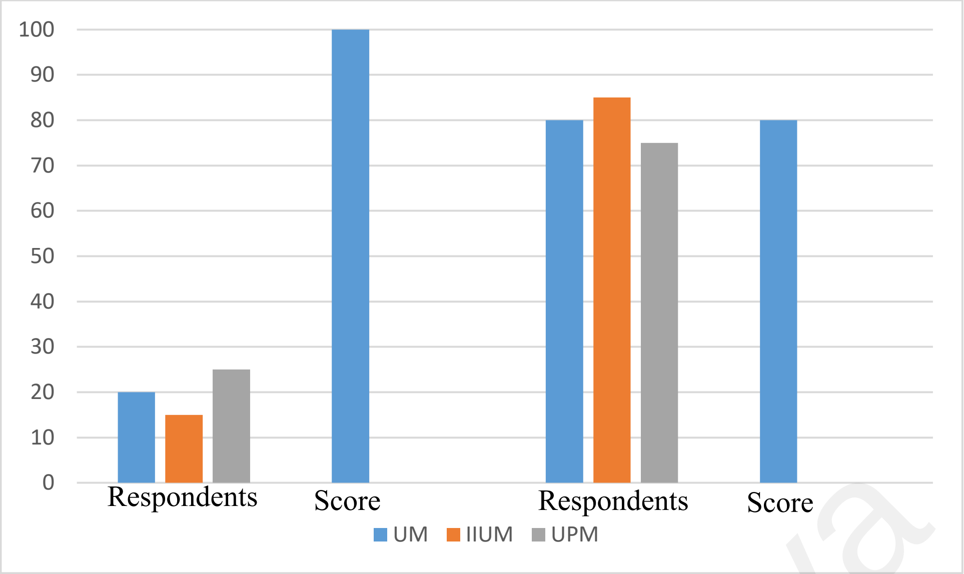


Figure 6.4: Comparative naturalness score based on university

The respondents from UM rated the system to be 84% in terms of intelligibility, the respondents from IIUM rated the system to be 81% in terms of intelligibility and respondents from UPM rated the system to be 82% in terms of intelligibility. Figure 6.4 depicts the comparative score of intelligibility based on university.

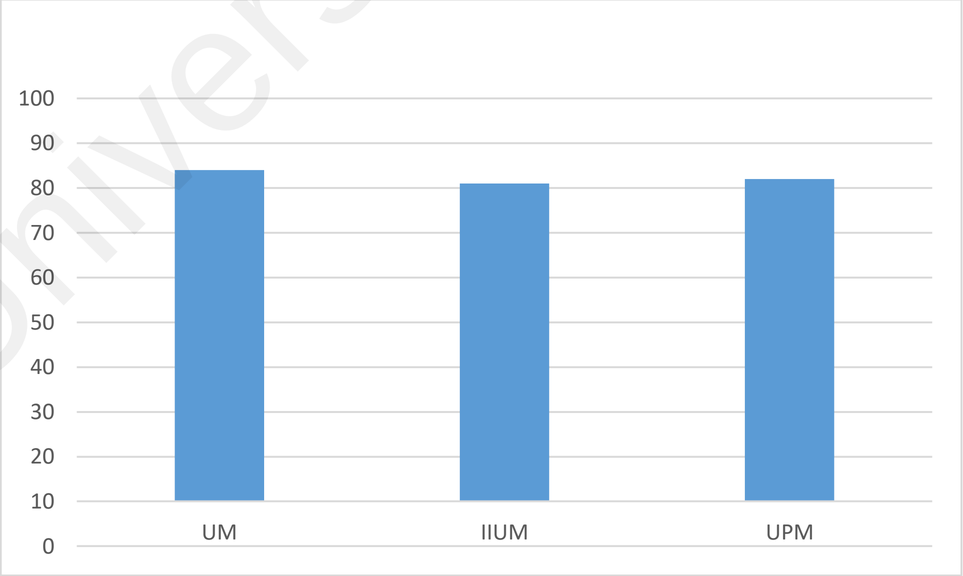


Figure 6.5: Comparative intelligibility score based on University

In term of gender, 8 out of the 33 male respondents and 2 out of the 17 female respondents rated the system to be 100% in terms of naturalness while the rest rated the system to be 80%. Figure 6.6 depicts the comparative naturalness score based on gender.

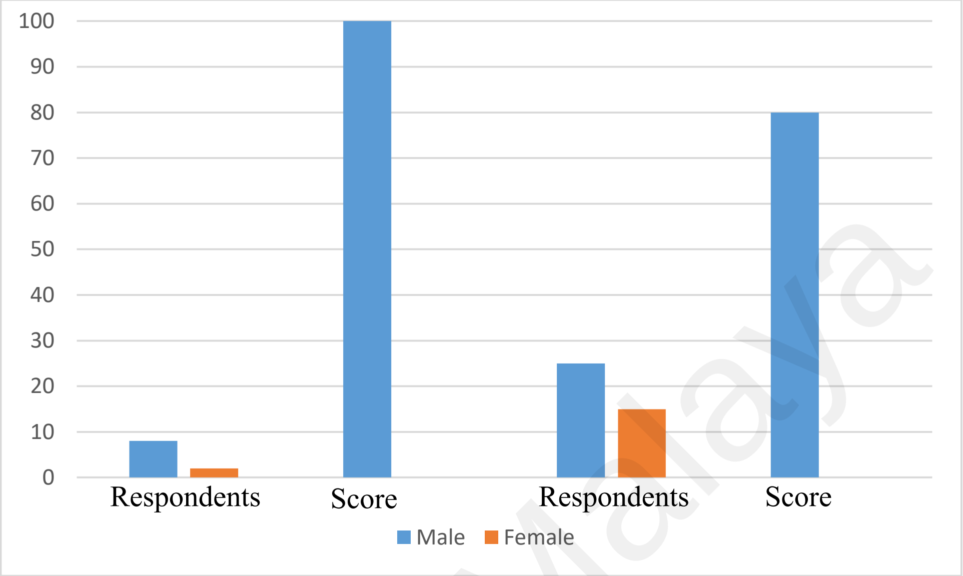


Figure 6.6: Comparative naturalness score based on gender

The overall intelligibility score for the male respondents was found to be 83.2 and that of female was 80.2 as illustrated in figure 6.7.

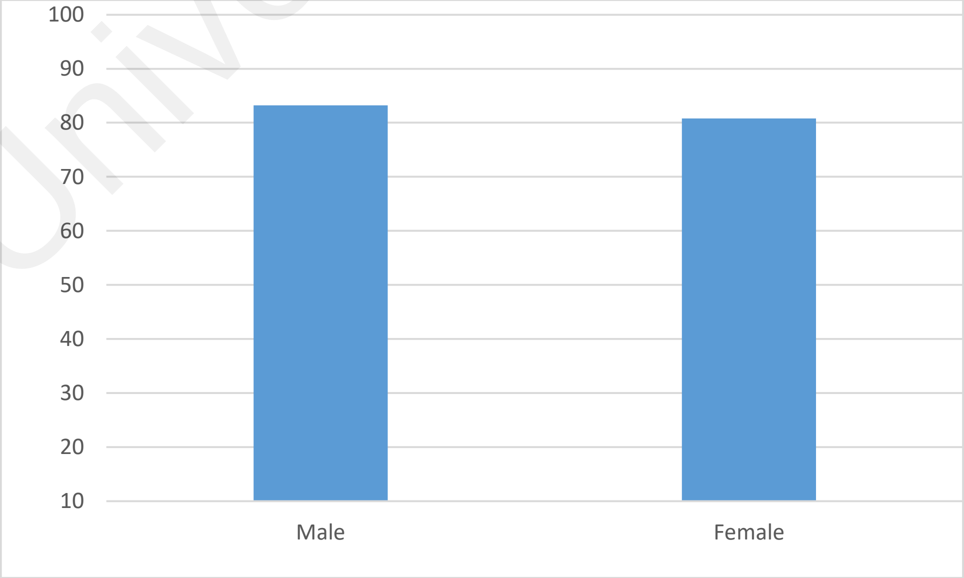


Figure 6.7: Comparative intelligibility score based on gender

At the end of the evaluation, it was found that the synthesized speeches was actually very easy to be comprehended with only few difficult words such as “watsi, zaaba, and gyara”. It also shows that the male respondents have a better comprehension of the synthesized than the female respondent, this might be as a result of the male voice used for the training of the system. Figure 6.8 depicts the evaluation session with some of the respondents.

Figure 6.8: Evaluation session with respondents from UM, IIUM and UPM

6.5 Chapter summary

In this chapter, the listening evaluation was conducted to determine the performance of the developed system in term of the intelligibility and naturalness. From the result, the cross-lingual technique using the DNN-based synthesis outperformed the existing HMM-based cross-lingual TTS system of other under-resourced languages by 10% in terms of naturalness and 5% in terms of intelligibility. There is no element of biasness among the respondents as only 4% of male evaluators gives a 5 upon 5 score though the synthetic speech is based on the male voices.

CHAPTER 7: CONCLUSION AND FUTURE WORK

This chapter summarizes the findings and the contributions of this research concerning the development of a TTS system for Hausa, an under resourced language. This chapter also revisits the objectives of this research and how they were achieved. On top of that, this chapter discusses the limitation as well as the possible future research direction.

The major factor that hinders the development of the TTS system for Hausa is that being it an under-resourced language, the language does not have the adequate resources required for the development of a TTS system. The resources required to develop a TTS system includes recorded speech, transcription, segmentation, and labeling. Although the DNN-based speech synthesis is effective for developing a TTS system for under resourced language using the existing resources of a well-resourced language, the system can only synthesize the neutral speech.

7.1 Research Objectives

The three objectives that have been fulfilled in this research are as follows:

7.1.1 Achieving Research Objective 1

The first research objective was to accumulate the resources required for the development of a TTS system for Hausa as an under-resourced language. To achieve this objective, the literature review was performed to gather the information on the existing TTS system for Hausa if any, as well as the existing resources for the development of a new TTS system. The review was also conducted to obtain the information on how the TTS system for under-resourced languages can be developed with minimal resources. This review of literature also point to the lack of resources needed for the development of TTS system for Hausa and also suggests how the resources of a well-resourced language can be used for developing the TTS system for Hausa.

Due to the absence of suitable resources for the development of a TTS system for Hausa, the key minimal resources for Hausa were accumulated in this research. The resources developed in this research include Hausa text corpus, Hausa speech corpus, pronunciation dictionary and the speech acoustic model of Hausa.

7.1.2 Achieving the Research Objective 2

The second objective of this research is to identify suitable state-of-the-art technique(s) for the development of TTS system for under-resourced languages. In this case, from the review, the most commonly used state-of-the-art techniques for the TTS system development of an under-resourced language is the HMM-based synthesizer with techniques such as the cross-lingual technique, which outperform the other techniques in the development of TTS system for under-resourced languages.

Though statistical parametric speech synthesis techniques include HMM and DNN, where the latter shows a better performance in term of intelligibility and naturalness, better noise robustness, and reduced WER, the DNN-based was not considered for the development of TTS system for under resourced language. For Hausa, the cross lingual adaptation makes use of the English language resources. This is because, both English and Hausa use the Roman alphabet as a writing system and also share many similarities in their segmental phonemes as illustrate in table 3.2 and table 3.3 in chapter 3.

7.2.3 Achieving the Research Objective 3

The third objective of this research is to develop and evaluate (in terms of intelligibility and naturalness) the TTS system for Hausa. The development of the TTS system makes use the cross-lingual technique and the DNN-based synthesis. Listening evaluation is conducted using 50 Hausa native listeners. The listening evaluation shows that the average naturalness is 84% (4.2 out of 5), while average intelligibility is 82% (4.1 out of

5). The performance of the DNN-based Hausa TTS system was better than the existing TTS system for under-resourced TTS system (Taiwanese, Tamil, and Indian Eng. Lan.) developed using the HMM-based speech synthesizer.

7.2 Research Contributions

The main contributions of this research are as follow:

(a) ***TTS system for Hausa Language***

The TTS system developed in this research could be the first TTS system for the Hausa language, which is a great advantage to the Hausa community and the world at large, as it will help not only the visually impaired but also those who want to learn Hausa as the second language. In human computer interaction, speech synthesis also helps people in general, such as application like google map, audio books, and transport scheduled reading system, which ease the efforts for people for interacting with this application.

(b) ***DNN-based TTS system for under-resourced language***

Although the DNN-based TTS system in some instances performed better than the HMM-based TTS system, at the present moment, it was yet to be developed for the under-resourced languages. This research is the first to experiment the development of TTS system for under-resourced language using the DNN. From the experiment conducted, it was found that the intelligility and naturalness of the proposed DNN-based TTS system was better than many of the HMM-based TTS system developed for under-resourced languages.

(c) ***Experiment the adaptability of cross-lingual approach to DNN-based synthesis***

This research has experiments the suitability of the cross-lingual technique to be applied to the DNN-based synthesis, proofing the adaptability of the cross lingual technique with improved performance. The finding from this research can help many of the under-resourced languages toward the development of a TTS system with high performance in terms of intelligibility and naturalness. Although the database in this research is limited in size and vocabulary, the use of the cross-lingual approach and the DNN-based synthesis overcome the resource limitation. As such this research can serve as a good guideline for future development.

7.3 Research Limitations

Some of the limitation of this research is that the developed Hausa TTS system can only synthesize the neutral speech and not emotional speech. On top of that, the system is only synthesizing the male speech. It is not known how the TTS system performs when synthesizing female speech. To enable the TTS system to synthesize the emotional speech, more work is needed such as the recording of the emotional speech. On top of that, the developed TTS system performs badly in synthesizing some glottalic consonant (e.g ƙ, ɓ, ɗ represented by k, b, d respectively). The limitation of vocabulary size, the use of only male voice, and small number of respondents for listening evaluation has influenced the finding in this research.

7.4 Future Research

The future direction of this research is to develop the acoustic model for different dialect, thus enabling the TTS system to pronounce the words from all the Hausa dialect. On top of that, there is need to build a larger speech database that can synthesize both the neutral and emotional speech for the under-resourced languages, which will no doubt increase the acceptability of the TTS system by the society.

7.5 Conclusion

This research described the adaptability of the cross-lingual approach to the DNN-based synthesis for the development of TTS system for under-resourced languages such as Hausa. In view of the resource scarcity for TTS system development, the development of Hausa begins with the accumulation of minimal resources such as text and speech corpus, transcription, segmentation and labeling, training, and testing. As the DNN-based TTS system is relatively new to under-resourced languages, listening evaluation was conducted to measure the performance of the TTS system in terms of naturalness and intelligibility.

Due to the limited recorded speech database, this research have experimented two forms of DNN training, which are the three hidden layers and two hidden layer to determine the best training. It was found that three hidden layers with a total of 100 neurons outperform the two hidden layer with a total of 75 neurons. Thus, it can be suggested that the speech acoustic model improved when the number of neurons is increased, increasing the naturalness and intelligibility of the synthesized speech.

REFERENCES

Adamu, M. (1978). *The Hausa Factor in West African History*: Ahmadu Bello University Press Zaria.

Adamu, M. (1984). The Hausa and their neighbours in the central Sudan. *Niane, ed*, 266-300.

Balyan, A., Agrawal, S., & Dev, A. (2013). *Speech synthesis: A review.* Paper presented at the International Journal of Engineering Research and Technology.

Berment, V. (2004). *Méthodes pour informatiser les langues et les groupes de langues «peu dotées».* Université Joseph-Fourier-Grenoble I.

Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech Communication, 56*, 85-100.

Biadsy, F., Habash, N., & Hirschberg, J. (2009). *Improving the Arabic pronunciation dictionary for phone and word recognition with linguistically-based pronunciation rules.* Paper presented at the Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics.

Black, A. W., & Taylor, P. (1994). *CHATR: a generic speech synthesis system.* Paper presented at the Proceedings of the 15th conference on Computational linguistics-Volume 2.

Boothalingam, R., Solomi, V. S., Gladston, A. R., Christina, S. L., Vijayalakshmi, P., Thangavelu, N., & Murthy, H. A. (2013). *Development and evaluation of unit selection and HMM-based speech synthesis systems for Tamil.* Paper presented at the National Conference on Communications (NCC2013).

Chomphan, S., & Kobayashi, T. (2008). Tone correctness improvement in speaker dependent HMM-based Thai speech synthesis. *Speech Communication, 50*(5), 392-404.

Clark, R. A., Podsiadlo, M., Fraser, M., Mayo, C., & King, S. (2007). *Statistical analysis of the Blizzard Challenge 2007 listening test results.* in Proceeding Blizzard Challenge Workshop 2007 (in Proc. SSW6), Bonn, Germany, Aug. 2007.

Esmeir, S., & Markovitch, S. (2007). Anytime learning of decision trees. *Journal of Machine Learning Research, 8*(5), 891-933.

Fan, Y., Qian, Y., Soong, F. K., & He, L. (2016). *Speaker and language factorization in DNN-based TTS synthesis.* Paper presented at the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2016).

Ferreira, J. P., Chesi, C., Baldewijns, D., Braga, D., Dias, M., & Correia, M. (2016). *The first Mirandese text-to-speech system*. Paper presentend at the Language Documentation and Conservation Special Publication, Jan., 2016, 150-158.

Figueiredo, A., Imbiriba, T., Bruckert, E., & Klautau, A. (2006). *Automatically estimating the input parameters of formant-based speech synthesizers.* Paper presented at the Workshop de Tecnologia da Informação e da Linguagem Humana-TIL 2006. October, 23-27.

Fraser, M., & King, S. (2007). *The blizzard challenge 2007*. In: Proc. Blizzard Challenge Workshop 2007 (Bonn, Germany), 2007.

Gutkin, A., Ha, L., Jansche, M., Kjartansson, O., Pipatsrisawat, K., & Sproat, R. (2016). Building Statistical Parametric Multi-speaker Synthesis for Bangladeshi Bangla. *Procedia Computer Science, 81*, 194-200.

Gutkin, A., Ha, L., Jansche, M., Pipatsrisawat, K., & Sproat, R. (2016). TTS for Low Resource Languages: A Bangla Synthesizer.

Hanzlíček, Z. (2010). *Czech HMM-based speech synthesis.* Paper presented at the International Conference on Text, Speech and Dialogue.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., . . . Sainath, T. N. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine, 29*(6), 82-97.

House, A. S., Williams, C. E., Hecker, M. H., & Kryter, K. D. (1965). Articulation-Testing Methods: Consonantal Differentiation with a Closed-Response Set. *The Journal of the Acoustical Society of America, 37*(1), 158-166.

Indumathi, A., & Chandra, E. (2012). Survey on speech synthesis. *Signal Processing: An International Journal (SPIJ), 6*(5), 140.

Isewon, I., Oyelade, J., & Oladipupo, O. (2014). Design and Implementation of Text To Speech Conversion for Visually Impaired People. *International Journal of Applied Information Systems, 7*(2), 25-30.

Iskarous, K., Goldstein, L., Whalen, D. H., Tiede, M., & Rubin, P. (2003). *CASY: The Haskins configurable articulatory synthesizer.* Paper presented at the International Congress of Phonetic Sciences, Barcelona, Spain.

Justin, T., Mihelič, F., & Žibert, J. (2016). Towards automatic cross-lingual acoustic modelling applied to HMM-based speech synthesis for under-resourced languages. *AUTOMATIKA: časopis za automatiku, mjerenje, elektroniku, računarstvo i komunikacije, 57*(1), 268-281.

Karjalainen, M. (1999). *Review of Speech Synthesis Technology.* Master's Thesis, Helsinki University of Technology., [online]. Available: http://www.acoustics.hut.fi/publications/files/theses/lemmetty_mst/thesis.pdf, last accessed: 22/02/2017

Kayte, S., & Gawali, B. (2015). A Text-To-Speech Synthesis for Marathi Language using Festival and Festvox. *European Journal of Computer Science and Information Technology, 3*(5), 30-41.

Keller, E. (1995). *Fundamentals of phonetic science.* Paper presented at the Fundamentals of speech synthesis and speech recognition, 7(2), 5-21.

Kim, Y.-J., & Conkie, A. (2002). *Automatic segmentation combining an HMM-based approach and spectral boundary correction.* Paper presented at the INTERSPEECH.

King, S. (2010). A beginners' guide to statistical parametric speech synthesis. Paper presented at the Centre for Speech Technology Research, University of Edinburg, UK.

Krauwer, S. (2003). The basic language resource kit (BLARK) as the first milestone for the language resources roadmap. *Proceedings of SPECOM 2003*, 8-15.

Le, V.-B., & Besacier, L. (2009). Automatic speech recognition for under-resourced languages: application to Vietnamese language. *IEEE Transactions on Audio, Speech, and Language Processing, 17*(8), 1471-1482.

Ling, Z.-H., Kang, S.-Y., Zen, H., Senior, A., Schuster, M., Qian, X.-J., . . . Deng, L. (2015). Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends. *IEEE Signal Processing Magazine, 32*(3), 35-52.

Maas, A. L., Qi, P., Xie, Z., Hannun, A. Y., Lengerich, C. T., Jurafsky, D., & Ng, A. Y. (2017). Building DNN acoustic models for large vocabulary speech recognition. *Computer Speech & Language, 41*, 195-213.

Maia, R., Zen, H., Tokuda, K., Kitamura, T., & Resende Jr, F. G. V. (2003). *Towards the development of a brazilian portuguese text-to-speech system based on HMM.* Paper presented at the INTERSPEECH, 2003.

Malah, Z., & Rashid, S. M. (2015). *Contrastive Analysis of the Segmental Phonemes of English and Hausa Languages*. Paper presented at the International Journal of Languages, Literature and Linguistics, Vol. 1, No. 2, June 2015.

Mandal, S. K. D., & Datta, A. K. (2007). *Epoch synchronous non-overlap-add (ESNOLA) method-based concatenative speech synthesis system for Bangla.* Paper presented at the SSW.

Molapo, B., Barnard, E., & De Wet, F. (2014). *Speech data collection in an under-resourced language within a multilingual context.* Paper presented at the 4th International Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU), 2014.

Mporas, I., Lazaridis, A., Ganchev, T., & Fakotakis, N. (2009). *Using hybrid HMM-based speech segmentation to improve synthetic speech quality.* Paper presented at the 13th Panhellenic Conference on Informatics, 2009. (PCI'09).

Mukherjee, S., & Mandal, S. K. D. (2014). A Bengali HMM based speech synthesis system. *arXiv preprint arXiv:1406.3915*.

Mullah, H. U., Pyrtuh, F., & Singh, L. J. (2015). *Development of an HMM-based speech synthesis system for Indian English language.* Paper presented at the 2015 International Symposium on Advanced Computing and Communication (ISACC 2015).

Mumtaz, M. B., Ainon, R. N., Roziati, Z., Don, Z. M., & Gerry, K. (2011). *A cross-lingual approach to the development of an HMM-based speech synthesis system for Malay*. Paper presented at ISCA. INTERSPEECH, 2011.

Mustafa, M. B., Don, Z. M., Ainon, R. N., Zainuddin, R., & Knowles, G. (2014). Developing an HMM-based speech synthesis system for Malay: a comparison of iterative and isolated unit training. *IEICE transactions on information and systems, 97*(5), 1273-1282.

Navas, E., Hernaez, I., & Luengo, I. (2006). An objective and subjective study of the role of semantics and prosodic features in building corpora for emotional TTS. *IEEE Transactions on Audio, Speech, and Language Processing, 14*(4), 1117-1127.

Onaolapo, J., Idachaba, F., Badejo, J., Odu, T., & Adu, O. (2014). *A Simplified Overview of Text-To-Speech Synthesis*. In: Proceedings of the World Congress on Engineering, July 2 - 4, 2014, London, U.K.

Palkar, S., Black, A. W., & Parlikar, A. (2012). *Text-To-Speech for languages without an orthography.* Paper presented at the COLING (Posters).

Pärssinen, K. (2007). *Multilingual text-to-speech system for mobile devices: Development and applications*. Thesis for the degree Doctor of Technology, presented at Tampere University of Technology, 2007, 17-35.

Philips, J. E. (2004). Hausa in the twentieth century: An overview. *Sudanic Africa, 15*, 55-84.

Potard, B., Aylett, M. P., Baude, D. A., & Motlicek, P. (2016). Idlak Tangle: An Open Source Kaldi Based Parametric Speech Synthesiser based on DNN. *Interspeech 2016*, 2293-2297.

Price, R., Iso, K.-i., & Shinoda, K. (2016). Wise teachers train better DNN acoustic models. *EURASIP Journal on Audio, Speech, and Music Processing, 2016*(1), 1-19.

Qian, Y., Soong, F., Chen, Y., & Chu, M. (2006). An HMM-based Mandarin Chinese text-to-speech system *Chinese Spoken Language Processing* (pp. 223-232): Springer.

Rashad, M., El-Bakry, H. M., Isma'il, I. R., & Mastorakis, N. (2010). *An overview of text-to-speech synthesis techniques.* Paper presented at the 4th international conference on communications and information technology, Corfu Island, Greece.

Rebai, I., & BenAyed, Y. (2015). Text-to-speech synthesis system with Arabic diacritic recognition system. *Computer Speech & Language, 34*(1), 43-60.

Sani, M. A. (1999). Tsarin Sauti Da Nahawun Hausa. *Ibadan: University Plc, pp. 61.*

Schlippe, T., Djomgang, E. G. K., Vu, N. T., Ochs, S., & Schultz, T. (2012). *Hausa large vocabulary continuous speech recognition.* Paper presented at the SLTU.

Schwarz, D. (2006). Concatenative sound synthesis: The early years. *Journal of New Music Research, 35*(1), 3-22.

Sharma, B., Adiga, N., & Prasanna, S. M. (2015). *Development of Assamese Text-to-speech synthesis system.* Paper presented at the 2015 IEEE Region 10 Conference (TENCON 2015).

Sharma, B., & Prasanna, S. M. (2016). Polyglot Speech Synthesis: A Review. *IETE Technical Review*, 1-24.

Sher, Y.-J., Chiu, Y.-H., Hsu, M.-C., & Chung, K.-C. (2010). *Develop a HMM-based Taiwanese text-to-speech system.* Paper presented at the 2010 2nd International Conference on Software Technology and Engineering (ICSTE).

Sitaram, S., Palkar, S., Chen, Y.-N., Parlikar, A., & Black, A. W. (2013). *Bootstrapping text-to-speech for speech processing in languages without an orthography.* Paper presented at the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing.

Stan, A., Yamagishi, J., King, S., & Aylett, M. (2011). The Romanian speech synthesis (RSS) corpus: Building a high quality HMM-based speech synthesis system using a high sampling rate. *Speech Communication, 53*(3), 442-450.

Stolcke, A., Mandal, A., & Shriberg, E. (2012). *Speaker recognition with region-constrained MLLR transforms.* Paper presented at the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2012).

Tamura, M., Masuko, T., Tokuda, K., & Kobayashi, T. (1998). *Speaker adaptation for HMM-based speech synthesis system using MLLR.* Paper presented at the the third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis.

Tokuda, K., Zen, H., & Black, A. W. (2002). *An HMM-based speech synthesis system applied to English.* Paper presented at the IEEE Speech Synthesis Workshop.

Van Niekerk, D. R., & Barnard, E. (2009). *Phonetic alignment for speech synthesis in under-resourced languages.* In proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech 2009), Brighton, UK: 880-883.

Voiers, W. D. (1977). Diagnostic evaluation of speech intelligibility. *Benchmark papers in acoustics, 11.*

Watts, O., Ronanki, S., Wu, Z., Raitio, T., & Suni, A. (2015). *The NST–GlottHMM entry to the Blizzard Challenge 2015.* Paper presented at the Proc. Blizzard Challenge Workshop, 2015.

Weiss, C., Oliveira, L. C., Paulo, S., Mendes, C., Figueira, L., Vala, M., . . . Andre, E. (2007). *eCIRCUS: building voices for autonomous speaking agents.* Paper presented at the SSW.

Wu, Z., Swietojanski, P., Veaux, C., Renals, S., & King, S. (2015). *A study of speaker adaptation for DNN-based speech synthesis.* Paper presented at the INTERSPEECH 2015.

Wu, Z., Watts, O., & King, S. (2016). Merlin: An open source neural network speech synthesis system. *Proc. SSW, Sunnyvale, USA*.

Yahaya, I. Y. (1988). *Hausa a rubuce: Tarihin rubuce rubuce cikin Hausa*: Kamfanin Buga Littattafai Na Nigeria Ta Arewa.

Yamagishi, J., Ling, Z., & King, S. (2008). Robustness of HMM-based speech synthesis.

Yamagishi, J., Nose, T., Zen, H., Ling, Z.-H., Toda, T., Tokuda, K., . . . Renals, S. (2009). Robust speaker-adaptive HMM-based text-to-speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing, 17*(6), 1208-1230.

Yamagishi, J., Onishi, K., Masuko, T., & Kobayashi, T. (2005). Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis. *IEICE transactions on information and systems, 88*(3), 502-509.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., & Kitamura, T. (1999). *Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis.* Paper presented at the Sixth European Conference on Speech Communication and Technology.

Zen, H. (2015). Acoustic modeling in statistical parametric speech synthesis–from HMM to LSTM-RNN. *In Proc. MLSLP, 2015*.

Zen, H., Senior, A., & Schuster, M. (2013). *Statistical parametric speech synthesis using deep neural networks.* Paper presented at the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013.

Zen, H., Tokuda, K., & Black, A. W. (2009). Statistical parametric speech synthesis. *Speech Communication, 51*(11), 1039-1064.