Taxonomy, Review and Research Challenges Of DNN-Based Text-To-Speech System for Hausa as Under-Resourced Language

Abubakar Ahmad Aliero[1], Dalhatu Muhammed[1] & Abubakar Ibrahim[2]
[1]Computer Science Department, Kebbi State University of Science and Technology, Aliero, Nigeria
[2] Asset Management Department (AMD), Nigeria Deposit Insurance Corporation (NDIC) , No. 15 Marina, Lagos , Nigeria

e-mail: abbatee4u@yahoo.com; dmaliero@yahoo.com; ibrahimab@ndic.gov.ng

Abstract - During the last few decades, researchers have continuously aims at improving the intelligibility and naturalness of Text-to-speech (TTS) system. Some of the major applications of TTS system includes document reader, speech translation, mobile read-aloud applications (such as google map reader), and announcement system. TTS system also serves as an assistive tool for disabled, which they use for reading online text, information, and automatic learning system for children. On top of that, TTS system is also a good way for preserving endangered languages due to globalization. Despite the potential benefits of TTS system, it was language dependent and was not developed for many of the languages around the world. One of the main issues is 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 approaches has 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 are only available for selected well-resourced languages like English, Arabic etc.

Index Terms - *Text –to-Speech, DNN, Hausa, Under –Resourced Language, SPSS, Cross-Lingual.*

1. INTRODUCTION

Hausa belongs to the Chadic family of languages of the Afro-Asiatic phylum and is spoken by more than 50 million people in West Africa as their mother tongue, second language and lingua franca. Hausa language is majorly spoken in the Sahel region of Africa, which consist of Northern Nigeria, Southern Niger, Southern Chad, Northern Cameron, Central Republic of African, and so on. Hausa is also spoken in some western countries like Germany, Hausa language has the highest number of speakers in Nigeria with over 29 million indigenous and 18 million non-indigenous speakers with majority of the indigenous speakers are from northern Nigeria and non-indigenous are from other neighbouring countries like Niger, Ghana, Cameroon, and Benin. Hausa language has been in existence since before the period of colonization, it was written in Arabic script,which is called the Ajami or hausa Arabic script [1]

Hausa consist 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 Hausa language.

Speech synthesis system or Text-to-speech (TTS) system is the process of generating human-like speech by computer from written text. Speech synthesis system enables people to conveniently listen to synthesized speech from various sources of written information with less effort and no longer have to stress their eyes while trying to read tiny fonts in those documents. Applications of speech synthesis system include call centre automation, automatic reading agent, chat avatar, intelligent robotics and game-based applications for improving the interaction between human and computers.
Speech synthesis has continued to be used in today's technology mainly to support people suffering from visual impairment and to those who are illiterate, where their difficulties in reading the written text has left a gap of successful communication. Speech synthesis system has greatly contributed in making people with these difficulties have access to many written resources [3].

TTS system plays a very important role in many aspect of life including learning, assistive tool, and many others. In leaning, TTS system can be used for language acquisition for children and adults. TTS system can be used as automatic learning system for children which will help them in learning words and their correct pronunciation. Speech Synthesis System for Hausa language will also be very useful for many applications such as GPS reader, public transit system, weather report, and so on.

TTS system can assist people with normal sight to listen to the synthesized speech from different version of written text without the use of much energy which also reduces the effect of reading directly from the screen, as nowadays people spend much of their time reading from the screen, be it reading from your mobile phone or from your computer screen, which courses eye discomfort, Headache, Difficult focusing and even burning sensations.

Under-resourced language refers to language that lack some or all of the resources required for the development of any speech technology system or application, such resources may include phonological system, orthography, linguistic expert, research in speech technology and so on. Many speech technology systems have been developed for several under-resourced languages with only little resources available.

The lack of progress in speech synthesis system for under-resourced languages is mainly attributed to non-availability of resources. Such as; recorded speech database, speech technology expertise, and funding issues. Lacks of speech technology expert, researcher, phonologist and machine learning experts have continued to put Hausa language as an under-resource language. Resources such as speech data, transcription, pronunciation dictionary, labels and letter-to-sound rules are the major problems for the development of TTS system for Hausa Language.

[2] Highlighted three major factors that affect the quality of synthesized speech: i. Vocoding, ii. Accuracy of acoustic model and iii. Over-smoothing. Which DNN have addresses the accuracy of acoustic model. Another research by Price et al., (2016) have shown that DNN have reduced the WER by 16.7% as compared to conventional HMM [4].

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 [5].

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

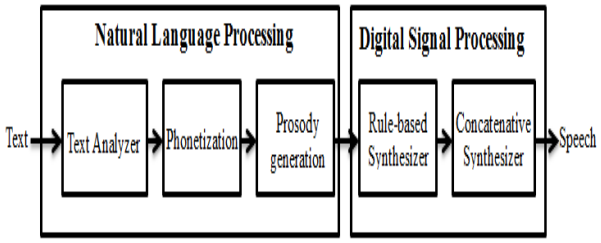

Figure Error! No text of specified style in document.**.1: Functional Diagram of Text-To-Speech System** [6].

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. 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 [7]. 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.

1.1.1.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 [7].

1.1.2.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.

1.1.3.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 [7].

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 [8].

The focus and scope of this study is to conduct a tentative review in order to come up with taxonomy of TTS techniques and outline the challenges and research direction for Hausa Language. The remaining section of these study are section 2 review of the related works, section 3 taxonomy of the speech synthesis techniques, section 4 TTS system for under resourced languages, section 5 Rapid development techniques for TTS system of under-resourced languages, section 6 open issues and research direction and section 7 conclusion.

2. REVIEW OF THE RELATED WORKS

This study 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.

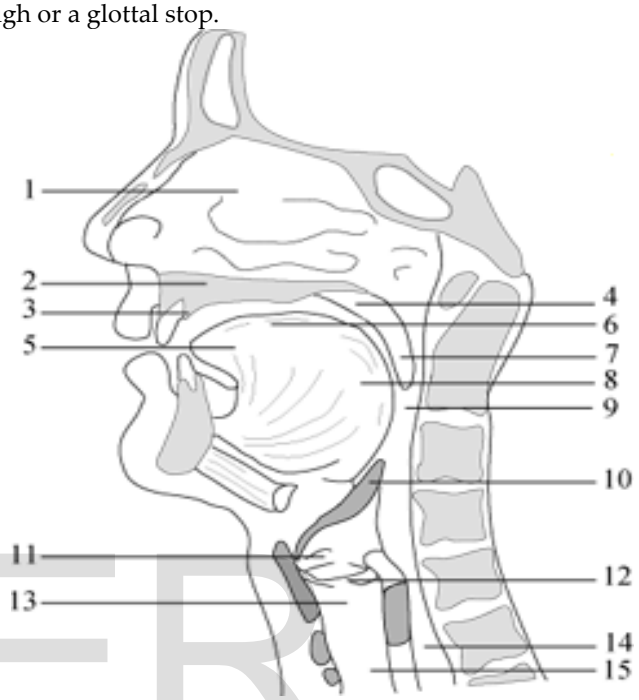

Figure **Error! No text of specified style in document.**.2: Human Speech Production System.
1. Nasal Cavity, 2. Hard palate, 3. Alveoral ridge, 4. Soft palate, 5. Tip of the tongue, 6. Dorsum, 7. Uvula, 8. Radix, 9. Pharynx, 10. Epiglottis, 11. False vocal cords, 12. Vocal cords, 13. Larynx, 14. Esophagus 15. Trachea [9].

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" [10]. 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.

3. TAXONOMY OF SPEECH SYNTHESIS TECHNIQUES

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 [6].

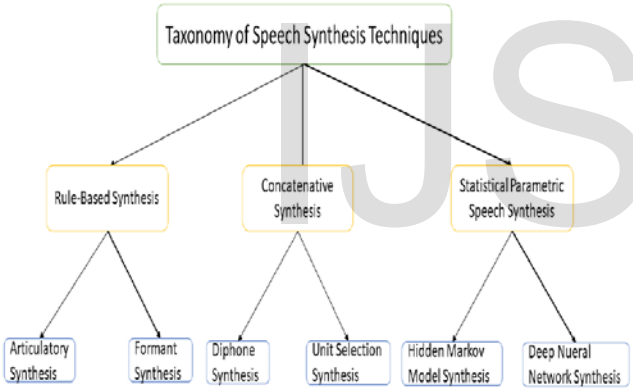

Figure 3.1: Taxonomy of Speech Synthesis Techniques

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.

3.1.1.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) [11] are not good as the formant synthesis or the concatenative synthesis [6].

3.1.2.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 [12].

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.) [12], 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 [6]. The formant synthesizer produces a monotonic and machine-like speech that clearly sounds unnatural.

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 [6].

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 [6]. 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.

3.2.1.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 [13]. For diphone synthesis, the quality of synthesized speech depends on the phonotactics of the language and the strength of the recorded speech [14]. 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.

3.2.2.Unit Selection Synthesis

A large speech database is used for the Unit Selection synthesis such as the ATR's CHATR [15]. 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 [5].
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 [5]. Another drawback of concatenative synthesis is that it can only synthesize speech units that are available in the database.

3.3. Statistical Parametric Speech Synthesis

Statistical Parametric Speech Synthesis (SPSS) based on the Hidden Markov Model [16] and Deep Neural Network [2] 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 [6].

3.3.1.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. [17].

Figure 3.2 shows the Architecture of HMM-based synthesis [17], 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.

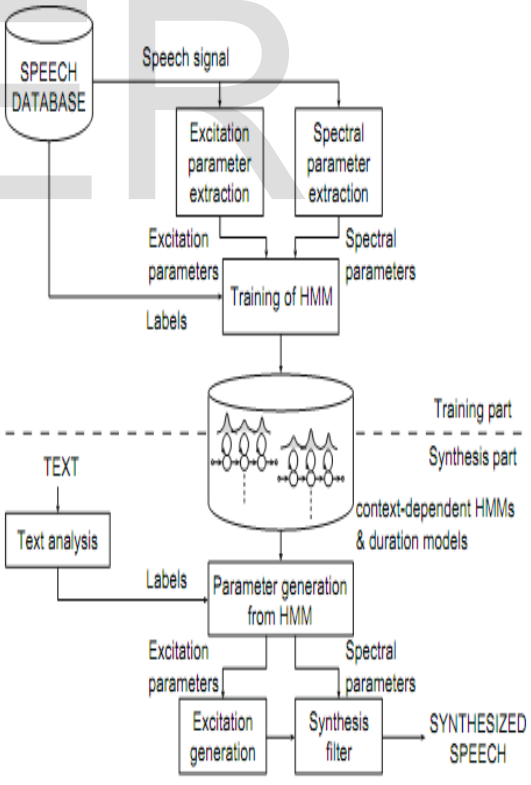

Figure 3.3: Architecture of HMM-based synthesis system [17].

The first language to pioneer the development of the HMM-based TTS system was the Japanese [18]. Table 3.1 shows the summary of the HMM-based TTS system developed for various languages. As shown in Table 3.1 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.

Author	Language	Language Status	Data Size	Training Data	Testing Data	Performance
[18]	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
[19]	Taiwanese	Under-resourced Language	Not mentioned	Not Mentioned	Not available	Performance = 4.0 of 5.0 (in terms of naturalness and intelligibility)
[20]	English language	Well-resourced language	Existing database (CMU communicator database)	524 sentences	Not mentioned	Very small runtime engine with less than 1MB
[21]	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
[22]	Czech	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

Table 3.1: Summary of Speech Synthesis System developed using HMM-based synthesis

3.3.2.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 [22].

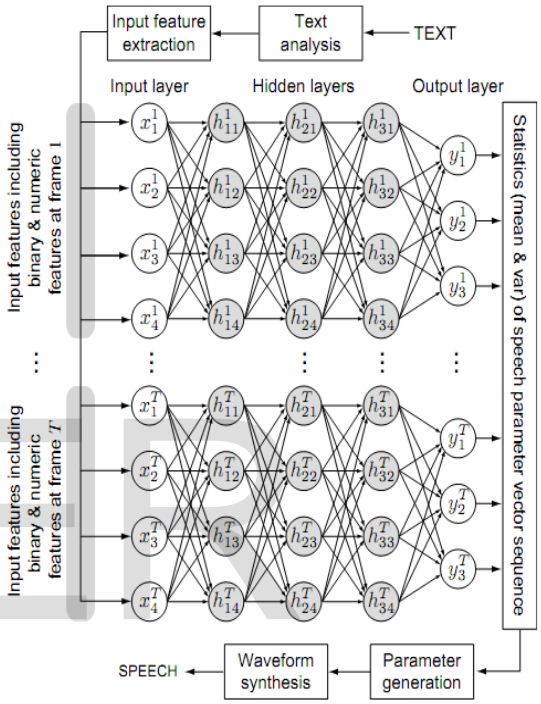

Figure 3.4: A speech synthesis framework based on DNN [2].

Figure 3.3 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 [23]. 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 [2].

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 [24].

Another advantage of SPSS such as HMM and DNN is its ability to change speaker characteristics, speaking style and emotion [17]. 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 [25].

Having succeeded with the Maximum Likelihood Linear Regression (MLLR) [25] and Maximum a posteriori (MAP) [22] 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 [25]. Table 3.2 provides the summary of speech synthesis system developed using the DNN-based technique. From Table 3.2, 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 3.2: Summary of TTS System Developed using DNN-based Synthesis

Author	Language	Data Size	Performance
[26]	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
[25]	English language	Voice Bank corpus (41 male and 55 female speakers)	better performance than the HMM baseline in terms of naturalness and speaker similarity
[17]	English language	CHiME-2 Corpus.	Performance = 6.7% better than the previous best result
[27]	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

Summary of the Speech Synthesis Techniques

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

Table 3.3: 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 obtainin data for articulator model. -Difficult in finding th balance between highl accurate model an easy to design an control
Formant Synthesis (Rule-based synthesis)	Generate speech using set of rules	Good in producing non-nasal and fricatives sound	-Produces machine lik speech which is clearl unnatural
Concatenative Synthesis	Generate speech by connecting natural pre-recorded speech units.	Good for synthesizing short units of words	-Only synthesiz phones that are defin within the speech uni inventory. -Requires large amoun of memory space.
Diphone Synthesis		Produces more natural speech than formant synthesizer	-It is discontinuos. -Not good for language with lot o inconsequence in th 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 amoun of database in speec recording. -The process o 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 b lower than the best Uni 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

4. 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 4.1 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 [30], used of well-resourced language resources for under-resourced TTS-system development [28], small storage footprint [29], rapid system development [30], 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.

Author	Language	Data Size	Training Data	Testing Data	Performance
[31]	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
[32]	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
[29]	Tamil	Not mentioned	Not mentioned	Not mentioned	Naturalness & Intelligibility = 3.86 out of 5.0
[33]	Mirandese	7 hours speech data	Not mentioned	Not mentioned	The system is relatively intelligible
[34]	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
[35]	Bengali	C-DAC corpus	816 sentences	10 sentences	The performance of system is 3.6 out of 5 in terms of intelligibility and naturalness
[36]	Bangladeshi Bangla	1891 utterances	Not mentioned	100 sentences	Server LSTM-RNN = 3.403 Embedded = 3.519 HMM = 3.43
[28]	Malay	1000 sentences	1000 sentences	50 utterances	Intelligibility Male = 99.33 Female = 99.11

Table 4.1: Summary of existing TTS systems developed for under-resourced languages

5. 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.

5.1. 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 5.1 below.

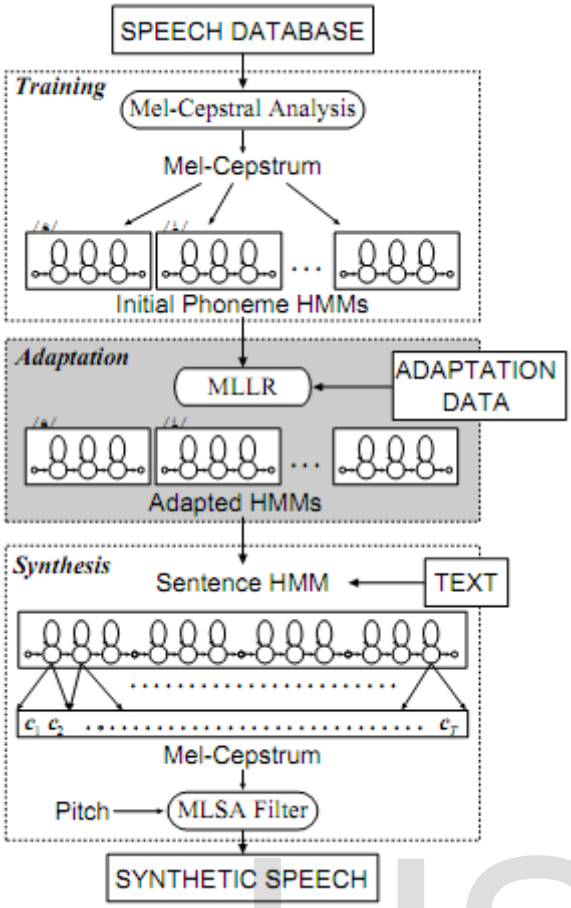

Figure 5.1: Block diagram of speech synthesis system using adaptation technique [38].

5.1.1.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.

5.1.2.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 [38].

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.

5.1.3.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 [39].

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 [28].

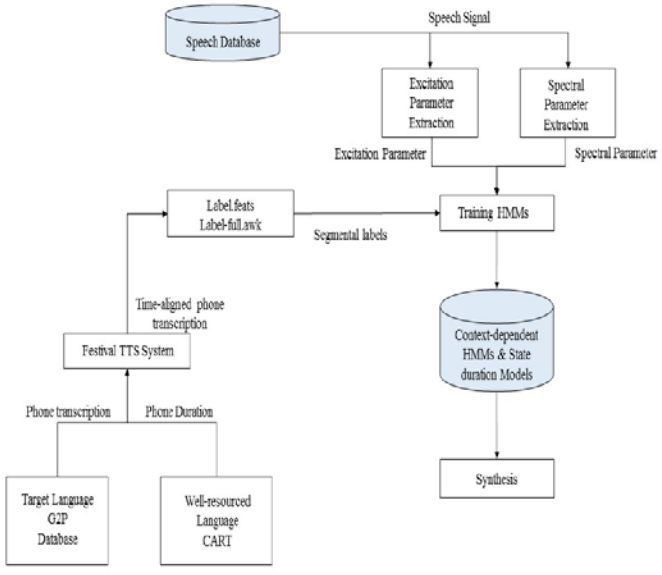

Figure 5.2: A Cross-lingual approach for development of TTS for under-resourced language [28]

In figure 5.2, 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 [36].

5.1.4.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.

5.1.5.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 [40].

5.2. 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 5.3 [41].

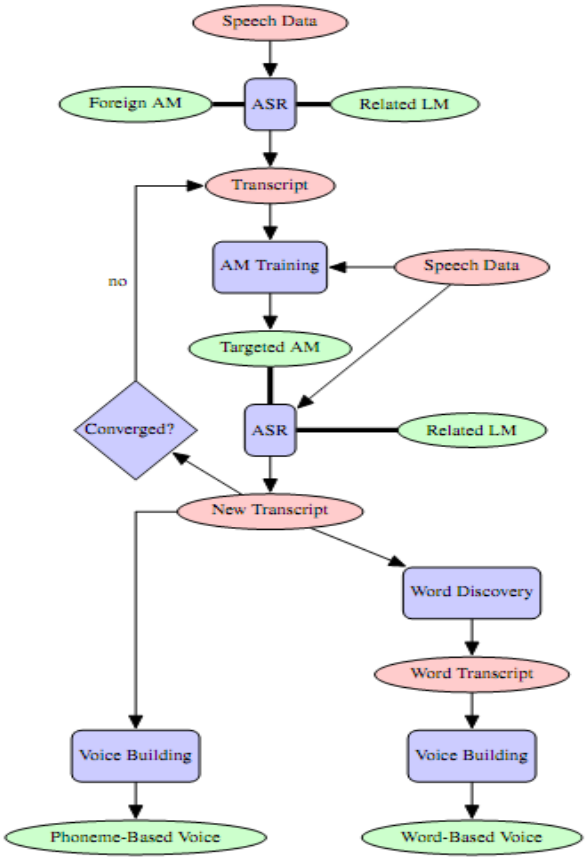

Figure 5.3: Bootstrapping Phonetic Transcription Technique [41].

6. OPEN ISSUES AND RESEARCH DIRECTION

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 [42]. 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 [43].

7. CONCLUSION

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 4.1 provides the summary for the development of TTS system for under-resourced languages. Table 4.1 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.

REFERENCES

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

[2] 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.

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

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

[9] 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

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

[11] 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.

[12] 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.

[13] 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.

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

[15] 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.

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

[17] 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.

[18] 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.

[19] 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).

[20] 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.

[21] 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.

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

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

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

[25] 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.

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

[27] 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).

[28] 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.

[29] 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).

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

[31] 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.

[32] 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).

[33] 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.

[34] 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.

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

[36] 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.

[37] 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.

[38] 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.

[39] 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).

[40] 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.

[41] 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.

[42] 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.

[43] 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.