Vol. 10(35), Jan. 2020, PP. 4460-4470

Contents list available at IJMEC

International Journal of Mechatronics, Electrical and Computer Technology (IJMEC)

Journal Homepage: www.aeuso.org

PISSN: 2411-6173 - EISSN: 2305-0543

A Cross-Lingual Text-To-Speech System for Hausa using DNN-Based Approach

Abubakar Ahmad Aliero[1], Dalhatu Muhammed[1], Mumtaz Begum Binti Peer Mustafa[2], Muhammad Saidu A[3]. and Muhammad Garba[1]

[1]*Computer Science Department, Kebbi State University of Science and Technology, Aliero, Kebbi, Nigeria*

[2]*Faculty of Computer Science & Information Technology, University of Malaya, Malaysia*

[3]*ICT Department, Kebbi State University of Science and Technology, Aliero, Kebbi, Nigeria*

***Corresponding Author's E-mail:** abbatee4u@yahoo.com*

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.

Keywords: *TEXT-TO-SPEECH, DNN, HAUSA, UNDER-RESOURCED LANGUAGE, CROSS-LINGUAL, SPSS.*

1. Introduction

Hausa language (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

Article History:
Received Date: Apr. 14, 2019
Accepted Date: Jun. 06, 2019
Available Online: Jan. 01, 2020

lingua franca. Hausa is spoken in the Sahel region of Africa, which consist of Northern Nigeria, Southern Niger, Southern Chad, Northern Cameron, and the Central Republic of African. Hausa is also spoken in some western countries like Germany, Hausa has the highest number of speakers in Nigeria with over 29 million indigenous and 18 million non-indigenous (neighboring countries like Niger, Ghana, Cameroon, and Benin). Hausa has existed before the period of colonization and was written in Arabic script called the Ajami [16].

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. TTS system serves as a document reader, speech translation, mobile read-aloud applications (such as Google map reader), and announcement system. It also serves as an assistive tool for disabled and is a way for preserving endangered languages due to globalization [7].

As TTS system was language dependent, it was not developed for many of the languages around the world. One of the main issues is the lack of the necessary resources, referred as the under-resourced language. Such resources include the phonological system, orthography, the linguistic expert, research in speech technology and so on. Many speech technology systems have been developed for several under-resourced languages with little resources available [5].

The lack of progress in TTS system for under-resourced languages is mainly attributed to non-availability of the resources such as the recorded speech database, speech technology expertise, and also funding issues. Hausa is classified as an under-resourced language due to lack of the necessary resources, which includes recorded speech database, transcription, pronunciation dictionary, labels, and letter-to-sound rules.

The state-of-the-art TTS systems are currently based on the Statistical Parametric Speech Synthesis system (SPSS) adopting models such as the Hidden Markov Model (HMM) [18]. There are three major factors that affect the quality of SPSS system in synthesizing speech, which are Vocoding, the accuracy of acoustic model and over-smoothing. These issues have been addressed by the Deep Neural Network (DNN)-based speech synthesis system [17]. It was shown in the existing work that DNN performs better than conventional HMM [17].

The major objective of this research work was to develop and evaluate a Text-To-Speech system for Hausa using the identified techniques, accumulated resources and evaluation methods, and in this case DNN and cross-lingual techniques were use for the development while listening method was used for the evaluation.

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.

2. Overview of TTS System for Under-resourced Languages

Over the last few decades 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 [14].

HMM-based synthesis has been used for the development of TTS system for under-resourced languages using the resources of a well-resourced language, which is formally known as the cross-lingual approach.

Maia, Zen, Tokuda, Kitamura, & Resende (2003), [10] have developed a TTS system for the Brazilian Portuguese language with a small amount of recorded database of about 600 sentences and only 200 phonetically balance sentences. A bootstrapping method was for training and a text-to-phoneme transcription was adopted using a phonetic transcriber. The TTS system generates synthetic speech with acceptable quality [10].

Mumtaz, Ainon, Roziati, Don, & Gerry (2011) [13] has developed a HMM-based TTS system for Malay language. 1,000 phonetically balanced sentences were created by crawling text from different sources, which was read by a native male and female speakers, with a total recording time of about four hours. A Grapheme-to-Phoneme rule was used in determining the phonemic representation of the words during the development of the TTS system. The system was developed using a cross-lingual approach by adopting English as the source language. The cross-lingual approach simplifies the development of TTS system for less-resourced languages. The developed system can produce a high-quality synthetic speech in terms of intelligibility and naturalness despite the use of small amount of database during the development [13].

In an effort to reduce the challenges for languages without orthography (i.e standard way of writing), Palkar, Black & Parlikar (2012), proposed an approach to developing a speech synthesis system for Marathi, a language without orthography. English and Hindi acoustic model is used for bootstrapping. Palkar, Black & Parlikar (2012), [15] build a speech database for Marathi and an Automatic Speech Recognition System of Hindi and English was used to generate the text from the input target corpora. The generated text was used in parallel as transcriptions for the two different Marathi TTS system [15]. On top of Hindi and English, resources of Telugu language was also considered. TTS system for Marathi developed using Telugu outperform the other two languages. The performance of the TTS system was dampened as the Telugu language has more phone set than the English language, forcing the different Telugu phone into same English phone.

In [5], a TTS system was developed for Bangla language using the HMM-based and Long Short-Term Memory Recurrent Neural Network (LSTM-RNN). Researchers proposed an efficient and faster way of bootstrapping a TTS system for under-resourced language by first using crowdsourcing in the development of speech corpus, and an existing text normalization system (Hindi language) was used to bootstrap the linguistic front-end for Bangla [5]. Three parametric synthesis systems were developed, two of which are based on LSTM-RNN method, and the other on is the classic HMM-based synthesis. All three were acceptable in terms of naturalness and intelligibility but the classic HMM- based TTS system outperform the LSTM-RNN. However, in terms of portability in handheld devices, the LSTM-RNN-based TTS system outperform the HMM-based TTS system.

Speech data acquisition is one of the most time-consuming in the development of TTS system for under-resourced languages. Justin, Mihelič & Žibert, (2016) [8] proposed an algorithm that can automatically extract similar acoustic of the target language from the source language, so as to simplify the development of TTS system for a new language. In the proposed Automatic phoneme mapping technique, the Slovene language was used as the target language, while the English Language is chosen as the well-resourced language.

In many cases, during the development of TTS system for under-resourced language a speech data has to be recorded, transcribed and labelled, which is time-consuming and also required more experience in the design. The proposed work in [8] has overcome this problem by automatically mapping the phoneme of the target language to that of the sourced language [8].

Table 1: TTS system developed for Under-Resourced Languages

Author/Year	Language	Technique	Size of Data	Merit	Weekness
Maia, Zen, Tokuda, Kitamura, & Resende (2003)	Brazilian Portuguese	HMM-based Text-to-Phoneme mapping	613 sentences	The method can determine the contextual informations and questions for decision tree-based context clustering.	For text processing it requires large amount of segmentation to increase the performance of the database.
Hanzlíček, (2010)	Czech	HMM-based STRAIGHT	5 hours speech data	Comparative with unit selection TTS system	Required large amount of data and takes long time for the training
Mumtaz, Ainon, Roziati, Don & Gerry (2011)	Malay	HMM-based cross-lingual approach	1000 senteces	Good interms of intelligibility	Poor naturality
Palkar, Black & Parlikar, (2012)	Marathi	HHM-based Bootstrapping/ speech recognition	30 minutes speech data	Good for languages without orthography	Uses many different resource languages to deteming a batter output.
Boothalingam, Solomi, Gladston, Christina, Vijayalakshmi, Thangavelu, & Murthy, (2013)	Tamil	HMM-Based	12 hours speech data	footprint-size is significantly smaller when compared to any of the FestVox-based voice	Very long time taken for the training and the output is poor interms of naturalness and intelligibility
Mukherjee & Mandal, (2014)	Bengali	HMM-Based phoneme mapping	816 sentences	training corpus used was actually meant for Automatic Speech Recognition	The intelligibility is very poor due to the type of speech database used for the training.
Kayte, & Gawali, (2015)	Marathi	di-phone concatenation approach		The system can convert a Unicode encoded Marathi text into human speech	The system lacks proper intonation modeling, therefore the speech generated is more like robotic Speech
Mullah, Pyrtuh, & Singh (2015)	Indian English	HMM-based	1000 Utterances	The method shows a good resulting run-time engine of HTS.	Naturallness = 3.0 of 5 likert scale Intelligibility = <4.0 of 5 likert scale

Gutki., Ha, Jansche, Kjartansson, Pipatsrisawat, & Sproat, (2016).	Bangladeshi Bangla	LSTM-RNN embedded HMM-based	1891 Utterances	Multiple speakers were used to collect data	Difficulty in training of the data because of the number of speakers used
Ferreira, Chesi, Baldewijns, Braga, Dias, & Correia, (2016)	Mirandese	HMM-based Graphene-to-Phoneme Mapping	7 hours speech data	Very good interms of intelligibility	The Naturity of the system is poor
Fan, Qian, Soong & He, (2016)	Mandarin	DNN-Based	900 sentences for 3 speakers each	Very good output interms of Naturallness and Intelligibility	It takes very long time in training of the data.
Justin, Mihelič, & Žibert, (2016)	Slovenian	HMM-based cross-lingual authoamtic phoneme mapping	2 min 21 sec. speech data	Good for manual phoneme mapping	Some phoneme of the domain language may be missing while mapping

From the above review it was observed that many TTS was developed for so many under-resourced languages using different techniques and methods with different size of database. These reviews also shows that cross-lingual approach provide a good output for a small speech data while the DNN-based thesis provides the best synthesized speech but requires large speech data for the training. Our aim is to use cross-lingual approach for the development of our system and train the data using DNN-based technique.

3. Proposed Development

Figure 1 illustrates a block diagram for the DNN-based speech synthesis system adapted from Fan, Qian, Soong, & He (2016), the diagram consists of both training and synthesis part. In training, the acoustic features for DNN output are first extracted from the speech signal with the feature extraction module, and the linguistic features for DNN input are to be generated by the proposed cross-lingual approach. The parameters of DNN are trained using pairs of input and output features with a mini-batched, back-propagation algorithm. During synthesis, the input text is first analyzed into labels, then mapped onto the acoustic features trained by DNN. In order to generate smooth parameter trajectories, 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.

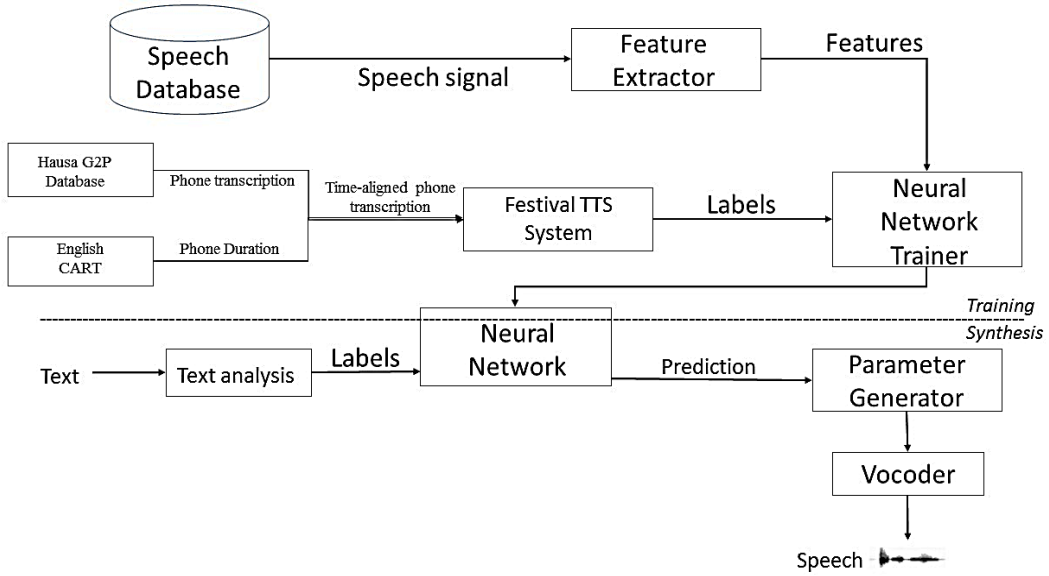

Fig. 1. Proposed architecture for the Hausa TTS system using cross-lingual approach and DNN adapted from Fan, Qian, Soong & He (2016).

3.1 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. 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. These tools, however, require an initial speech acoustic model of a particular language to perform the automatic alignment. 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 2 depicts the sample of segmentation and labeling using Praat software.

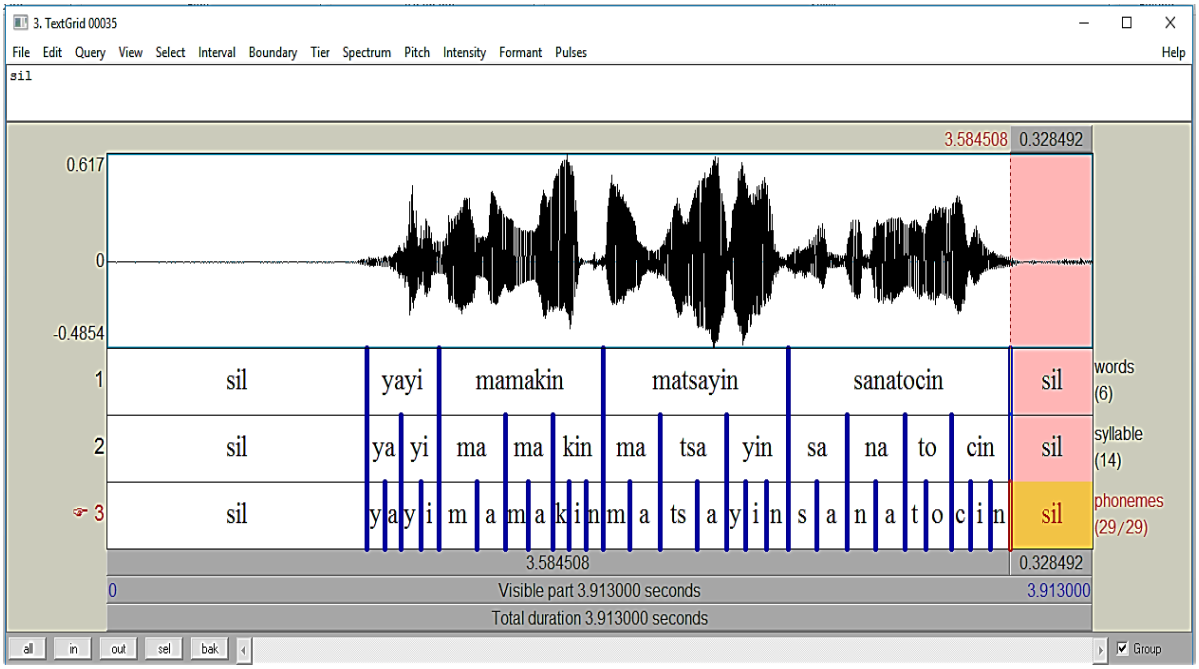

Fig. 2 Sample segmentation and labelling using Praat software "Yayi mamakin matsayin sanatocin/He was surprise with senate decision"

4. Result and Discussion

Building a rich speech corpus requires adequate phonetically rich and balanced sentences from the target language, phonetically rich and balanced sentences mean sentences that cover all phonemes of the target language and has adequately been represented in the complete 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 other short stories. A total of 179 sentences were created that has a good mixture of Hausa words, syllables, and phones. These sentences are within the range of five to seven words (29% of the sentences consist of 7 words, 26.3% consists of sentences with 6 words and 44.7% consists of 5 words). The sentences have a total of 1,046 words, 3,104 syllables and 7,325 total numbers of phonemes.

Our data consist of phonemes level transcription without time-alignment, speech data for each phoneme is transcribed in plain orthography at the utterance level. The input text was analyzed and normalized to generate a phonetic and contextual representation using a slightly modified text processing module of Kaldi for DNN training. This data was then time-aligned with the phone duration extracted from the English festival using dynamic time warping algorithms to generate the time-aligned phone transcription.

The data used for training the DNN acoustic model consist of about 50 utterances and 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 feedforward backpropagation neural network consists of three hidden layers each consists 50, 25, 25 units respectively.

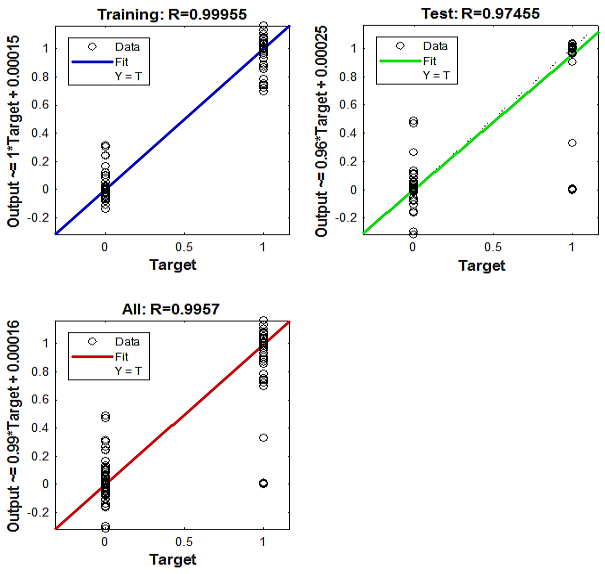

Fig. 3 Performance of the training at 233 epochs

The overall performance of the training using the three hidden layers is 99.57% which is obtained at 233 epochs.

5. Evaluation

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 [10] (at 20 % WER). During the listening test, it was found that most of the listeners can fully comprehend the synthesized speech after listening to the synthetic speech for one or two 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 4 depicts the result of the listening evaluation of the Hausa TTS system.

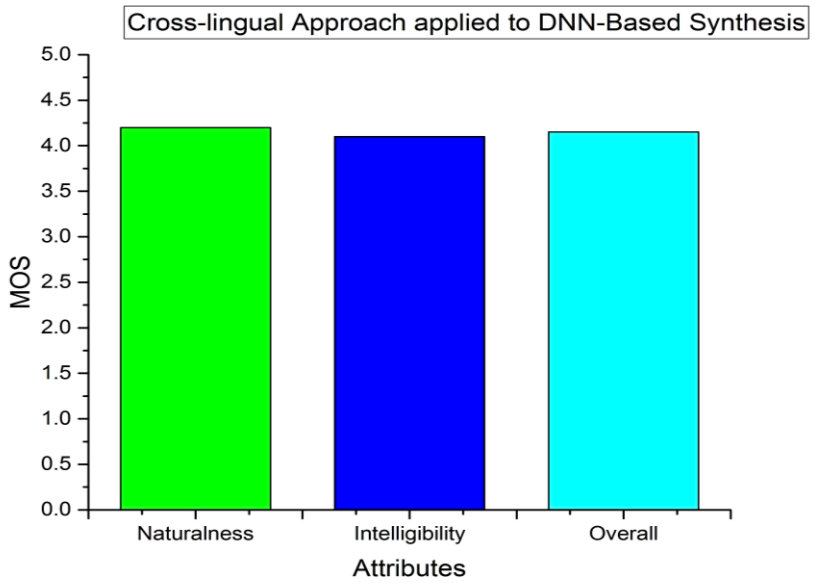

Fig. 4 Evaluation performance of Hausa TTS system

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 5 depicts the comparative naturalness score based on gender.

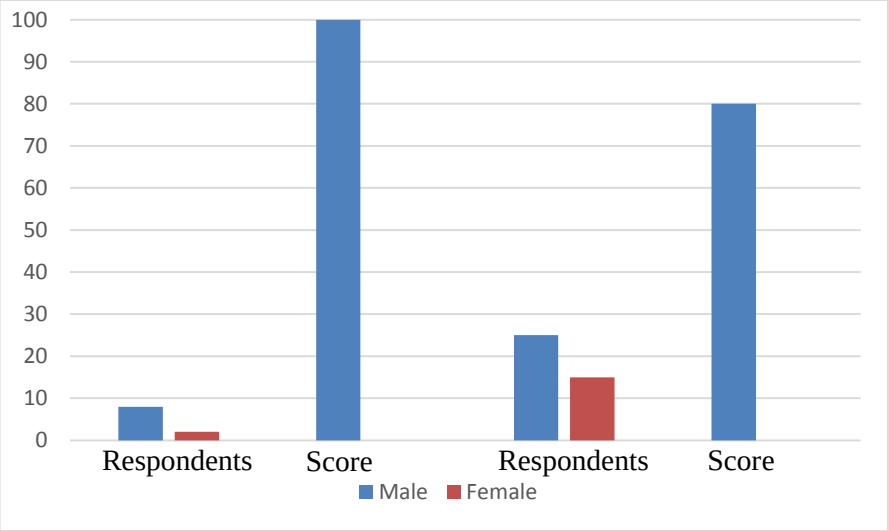

Figure 5: Comparative naturalness score based on gender

6. Comparative Analysis

Figure 6 and 7 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. 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).

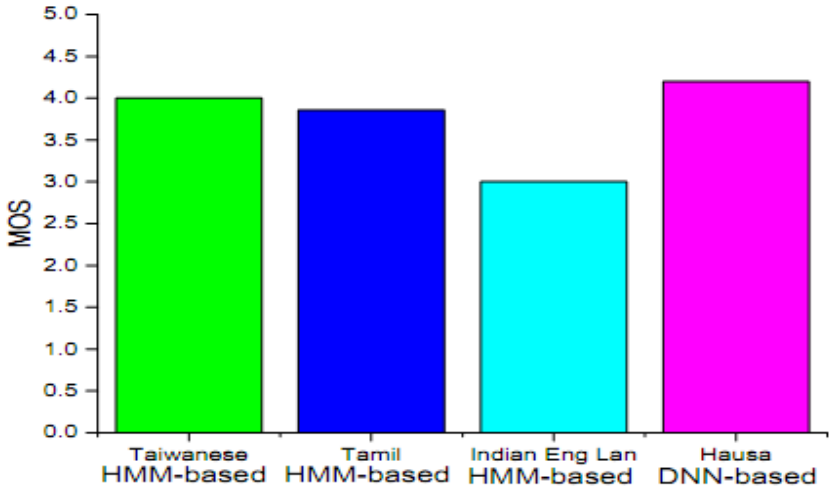

Fig. 6 Comparative score in terms of naturalness between HMM-based and DNN-based TTS systems developed for under-resourced languages.

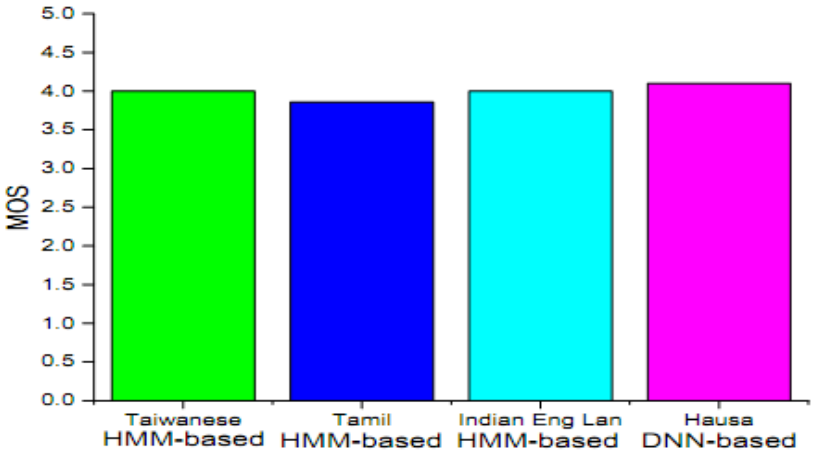

Fig. 7 Comparative score in terms of Intelligibility between HMM-based and DNN-based TTS systems developed for under-resourced languages.

Conclusions

Our proposed system have shown a remarkable improvement over the previous cross-lingual approaches that uses the HMM-based synthesis system for the development of TTS system for under-resourced languages. The experiment conducted in this study shows that the DNN-based synthesis produces better synthetic speech performance (in term of naturalness and intelligibility) than HMM-based synthesis using small data for training the speech acoustic model. 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. 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. This research also 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.

References

[1] 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 (NCC), 2013.

[2] 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 (ICASSP), 2016.

[3] Ferreira, J. P., Chesi, C., Baldewijns, D., Braga, D., Dias, M., & Correia, M. (2016). The first Mirandese text-to-speech system.

[4] 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.

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

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

[12] 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 International Symposium on Advanced Computing and Communication (ISACC), 2015.

[13] 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*.

[14] Okafor, K. C., Obayi, I. A. A., Ugwoke F.N., Ikechukwu V. O., "NigLT Ver 1.0: An Indigenous Language Translator Application for Major Ethnic Groups in Nigeria", International Journal of Advanced Scientific and Technical Research, India. Issue 4 Vol. 4, ISSN: 2249-9954, ISSN 2249-9954, 2013, Pp.882-890

[15] Palkar, S., Black, A. W., & Parlikar, A. (2012). *Text-To-Speech for Languages without an Orthography.* Paper presented at the COLING (Posters).

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

[17] 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.

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