TDNN-LSTM-based Acoustic Modeling for Verification of Qur’an Recitation for Non-Arabic Speakers using Kaldi Toolkit

Nazik O’mar Balula
College of Computer Science
and Information Technology
Sudan University of
Science and Technology

Mohsen Rashwan
Faculty of Engineering
College of Electronics and Communications
University of Cairo,Giza, Egypt

Sherif Mahdi Abdo
Faculty of Engineering
College of Electronics and Communications
University of Cairo, Giza, Egypt

ABSTRACT

Automatic Speech Recognition (ASR) has become an important component in HCI (Human -Computer Interaction) such as learning and processing natural languages. This paper provides a hybrid system which used GMM-HMM (Hidden Markov Model with a Mixture of Gaussians Model) and TDNN-LSTM (Time Delay Neural Network with Long-Short Term Memory Neural Network) to detect and correct the pronunciation errors in Qur’an recitation for non-Arabic speakers, specifically Indian speakers. The developed hybrid system concentrated on the ten Arabic letters (خ ح، ع، غ ، ض ، ظ ، ط ، د ، ق ، ص) that non-Arabic speakers can not pronounce them correctly and may confused with other letters that share the same articulation point. Training and Testing data collected from 94 Indian speakers MFCCs had been used as a feature extraction technique whereas GMM -HMM and TDNN-LSTM used as recognition tool. The main contribuation of the system is the enhancement and increament of accuracy of the HAFSS© system by using Deep Neural Network instead of GMM-HMM. The open-source Kaldi ASR toolkit recipes were used for building, training, testing and evaluation of the system. The developed system outperforms the GMM-MM model by 1.56% based on Kaldi toolkit word accuracy equation. The SUD (ص) letter accuracy using DNN-HMM model based on Kaldi toolkit outperforms the GMM-HMM model by 1% and at the same time outperforms DNN-HMM model based on HTK toolkit by 9.5%. The system acuracy was 95.14% using GMM-HMM and 96.88% using TDNN-LSTM.

Calculating the accuracy of the 10 Arabic letters, the best accuracy was 97.3% which achived by the letter TTA (ط), and the worest accuracy was 90.1% which achived by the letter DAA (د).
The rest of the paper is divided into seven parts, Section 1, Introduction introduced along with Qur’an recitation problems and Previous and Related studies. Section 2 outlines the Project Goal and Section 3 explains the structure of the system and Acoustic Model training steps. The acoustic model results explained in Section 4. Section 5 shows the Experiments Results and discussion along with Models Results comparison with previously published results . Conclusion is showed in Section 6 and Refrences in Section 7.

Gneral Tterms

Artificial intelligent, Speech Recognition,Deep Neural Network
keywords
Automatic Speech Recognition, Deep Neural network, kaldi toolkit, Time Delay neural network, Qur’anic recitation problems

1. INTRODUCTION

Automatic speech recognition (ASR) is the techneque that can identify words in spoken utterance and output them as text format that can be identified and readable by machine [25], it uses to build and develope automatic systems that can solve many problems in many fields such as NLP (Natural Languages Processing), voice and speech recognition and verification. It is very important technology that facilitates the interaction and communication between humans and computers.
ASR is still a challenging task due to the high viability in speech signals. For example, speakers may have different accents, dialects, or pronunciations, speak in different styles, at different rates and in different conditions.
The most important applications of ASR systems are the Qur’anic applications such as verification of Qur’an recitation and Qur’an memorization for all Muslims (Arabs and non-Arabs). To learn, read and recite Qur’an as reveald from Allah SWT, Tajweed rules (rules governing pronunciation during recitation) must be applied. There are still a lot of Muslims who can’t read and recite Qur’an

properly due to their unknowledge or little knowledge of Tajweed rules. Appling Tajweed rules ensure correct recitation and give the correct and intended meaning of the verses. Moreover, the correct recitation of Qur'an is required for all Muslims and indispensable in Islamic worship, such as prayers which is performed at least five times a day [14]. So that, errors and mistakes during recitation must be avoided . This issue leads to think of developing Computer-Aided Pronunciation Learning System (CAPL) to learn the correct way to recite Qur'an by pronounced letters correctly and applying the Tajweed rules to correct any mistakes in recitation without teacher, to help Muslims to memorize Al –Qur'an verses and verify their recitation properly and correctly in easy and fast way by themselves and to overcome the limitation of individual learning and lack of Qur'anic teachers who may not be available in some places and at a suitable time . ASR systems can facilitate learning,memorizing and verfing Qur'an recitation for many things.

1.1 Qur'an RECITATION PROBLEMS

Indiviual sounds are differ among people. The recitation of the same Qur'anic verse may be differ from one speaker to anorther. Although those verses were taken from the same Sura (chapter) due to differences in "Qraat" (Hafs, Kaloun, Warsh. . . etc) that used by reciters, or differences between written Qur'anic words and recitation of Qur'anic words due to appling the recitation rules. Also the consonants/vowel combinations and the co-articulation effect of emphatics and pharyngeals, pronounciation, Tanween and Ghonna rules and rules for combining words must be considered [**?**].

There is additional problem with non-Arabs, is that their pronunciation of Arabic letters is not proper as Arabs due to their accents, and this problem leads the user to fall in some mistakes when reading Qur'an verses. Wong utterance of words (pronunciation errors) and misreading words are the most errors that recitors falling on, therefore the system focused on detecting the non-proper pronunciation of uttered letters. The techniques used to verify and deal with non-Arabs (Indian speakers) recitation problems, was Deep learning techniques with Hidden Markov Model (TDNNs and HMMs) based on kaldi toolkit.

1.2 Previous and Related studies

Many speech recognition techniques for verification of Qur'an recitation have been introduced by many researchers. Here is some related works of verification of Qur'an recitation systems that introduced to overcome challenges facing Qur'an recitation. However, they do not attempt to cover the verification of Qur'an recitation for Arabs and non-Arabs to all Tajweed rules, which is an important issue for reciting Qur'an in proper way according to Tajweed rules. Some of difficulties in the traditional teaching of Qur'an recitation (face to face teaching) are mentioned by [7] such as :

1. One teacher teaches many students, so he/she cannot care to every student.
2. Students find difficulties in asking teacher due to shyness, hesitation and fear.
3. Teachers don't have full information about students background.
4. Difficulties and time shortness in accessing Tajweed books.
5. Material is delivered with same teaching method for all students irrespective of various attitudes of understanding.
6. Not all learners understand with the same style of teaching, some are visual learners, others are audio learners.
7. Non availability of teachers in every time and everywhere.

Due to all these problems researchers in the field of Automatic speech recognition systems introduced many available automated systems that help users to verify their recitation in easy and faster way in any place and at any time. A study done by [7], Tutoring system, helps in teaching and learning Tajweed rules, the main objective of this study is to overcome the difficulties faced in learning Tajweed, this system covers the first level of Tajweed with Rewaya Hafs from 'Aasem'. The system tested by 2 groups of users, students and teachers, these groups reported that the system overcome most problems of traditional teaching, the average satisfaction of teachers and students was 94.5% and 94% respectively.

The work that done by H. Tabbal [26], an Automated delimiter, which extracts verses from the audio files, this research taking into account the special ways to recite Qur'an with Tajweed rules, the goal of this thesis was to achieve new approach that uses speech recognition techniques to find and delimit verses in audio recitations automatically regardless of the reciter. The study use the Sphinx Framework as a research environment which is based on HMMs and used the SphinxTrain as a tool to develop the acoustic models. The feature extracted to the system by using MFCC algorithm, the Sphinx engine and acoustic models were used for recognition process.

Study done by [5], Computer Aided Pronunciation Learning (CAPL) system HAFSS© for teaching Arabic pronunciations to non-native speakers and also to teach correct recitation of Qur'an, this system is a helpfully system because it gives the user a suitable feedback of his/her errors in recitation if found, in addition to how can correct that errors . The HAFSS© system architecture consist of :

1. Verification HMM models : Is the acoustic HMM models for the system.
2. Speaker Adaptation : Is used to adapt acoustic models to each user acoustic properties in order to boost system performance.
3. Pronunciation hypotheses generator : This generates all possible pronunciation variants to test them against the spoken utterance.
4. Confidence Score Analysis : Analyzes the scores of the best decoded word sequence to determine the result.
5. Phoneme duration analysis : Recitation Rate Normalization (RRN) algorithm developed to overcome recitation speed variability by using phone duration to determine if phonemes have correct lengths or not.
6. Feedback Generator : produce useful feedback messages to the user by analyzing results from the speech recognizer. A database of 663 rules of pronunciation errors was connected with 2 other database to generate the suitable feedback (readable feedback or audible feedback) and to connect the recitation error with correct recitation rule. The system evaluated according to system judgment accept correctly pronounced phones or report same pronunciation error as the human expert.

   The below diagram demonstrate the architecture of HAFSS©.

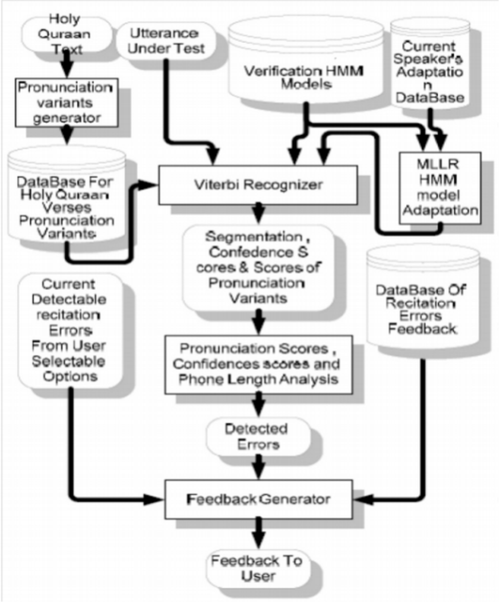

Holy Quraan Text
Utterance Under Test
Verification HMM Models
Current Speaker's Adaptatio n DataBase
Pronunciation variants generator
DataBase For Holy Quraan Verses Pronunciation Variants
Viterbi Recognizer
MLLR HMM model Adaptation
Segmentation , Confedence S cores & Scores of Pronunciation Variants
Current Detectable recitation Errors From User Selectable Options
DataBase Of Recitation Errors Feedback
Pronunciation Scores , Confidences scores and Phone Length Analysis
Detected Errors
Feedback Generator
Feedback To User

FIG. 1 – HAFSS© Block digram

Another work done by S. Hamid [17] who develop an automatic speech recognition system by implementing Computer-Aided Pronunciation Learning (CAPL) System. This system used many algorithms to detect and cover all user mistakes in recitation and gives a feedback to the user by his/her mistakes and the type of that mistakes and also give him/her the correct recitation. A Recitation Rate Normalization (RRN) algorithm was used to overcome the variability in recitation speed which may mislead the phone duration classification module, and HMM-based acoustic model speech recognition engine was implemented to detect the types of recitation mistakes and to segment input utterance, this system reducing the syllable error rate less than 5% for more than 50% of the speakers. The system was developed to Arabic language speakers only, and those who have previous knowledge about recitation rules. Another related work done by Sherif, M. A., Samir, A., Khalil, A.H. and Mohsen, R., CAPL for Qur'an recitation learning [23], which introduced to enhance the (CAPL) system HAFSS© which was developed for teaching Holy Qur'an recitation rules and Arabic pronunciations to non-native speakers, the verification done by using HMMs, the MLLR techniques used to increment the system performance by adapting the acoustic models.

The system was time consuming process because the baseline system used all data collected from certain user to create the new transformation. M. S. Abdo, A. H. Kandil, A. M. El-Bialy, S. A. Fawzy [4], introduced system to enhance usability of CAPL system for Qur'an Recitation Learning focused on detection of the non-proper pronunciation of a chosen set of uttered letters, new approach developed (MFCC) for automatic segmentation of the phonetic unit, also two algorithms applied, the first for verification of the speech phonetic unit that have an uttered recitation rules and the second was for automatically detecting the phonetic unit from the input utterance. The system accuracy was 100% in distinguishing between correct recitation and predictable errors.

The research that done by Hamid, S. E., and Rashwan, M. [11], which was introduced for Automatic Diagnosis of Pronunciation Errors, was used CAPL system with HMMs to generate the most probable pronunciation error hypotheses to test them against the spoken utterance from the user. HMMs was used as a detection tool for errors types. The system achieved good accuracy and less error rate. Focusing on Arabic speakers only was the limitation of this system. Automated Tajweed Checking Rules Engine for developing Qur'anic recitation with Tajweed rulse was system introduced by N. J. Ibrahim [13], to support Qur'anic learning process in effective and attractive way. This engine implemented and tested with the j-QAF students at primary school in Malaysia. MFCC algorithm used as feature extraction technique and HMMs used as a classifier. The speech samples were collected from 5 different reciters and saved as (.wav) files. The input of the system was the speech signal and phonetic transcription of the speech utterance. The recognition rate was 91.95% (ayates) and 86.41% (phonemes), after been tested on Sourate Al-Fatihah. The advantages of this engine is that it can help in learning and reciting Qur'an in a proper way without a teacher. However this engine was focused on one Qur'anic chapter (Sourate Al-Fatihah).

The proposed study by Ismail, A., Idris, M. Y. I., Noor, N. M., Razak, Z., & Yusoff, Z [15], for speech recognition, which introduced Checking Tool for Tajweed which concentrated on the Qalqalah Kubrah and Sughrah Tajweed rule for each(، ق ، ج ، ب ، د ، ط) Qalqalah letters by using hybrid algorithm Mel-Frequency Cepstral Coefficient and Vector Quantization (MFCC-VQ). The MFCC has been used as feature extraction techniques that convert voice signals to acoustic feature vectors and Vector Quantization (VQ) used as data reduction technique to reduce the data and this leads to speed up the system by reducing the computational time. The dataset collected by recording recitation from 45 reciters in three categories of reciters which are 20 males, 20 females and 5 children. The study observed that the speed performance of the hybrid algorithm MFCC-VQ is better than conventional MFCC by 86.928% for male, 94.495% for female and 64.683% for children. The study compared between hybrid algorithm MFCC-VQ and conventional MFCC and conclude that the MFCC-VQ was better in term of speed performance. However the study focused on Sourate Al-Ikhlas and Qalqalah phoneme only.

A new application of recitation verification of Qur'an based on correct makhraj, introduced by A. Wahidah and M. Suriazalmi [27], as a new way to learn reciting Qur'an in proper way and to reduce the duration time of learning from the expert. To obtain the correct makhraj the system used combination of the sound of hijaiyah letter (there is 29 basic Hijaiyah letter used in the Holy Qur'an) as the input data. The input speech was taken from people who are expert in makhraj utterance between the ages of 21 and 23, voices has been saved in (.wav) files by using Audacity Version 1.3 Software . Mel Frequency Cepstrum Coefficient (MFCC) used as feature extraction technique and Mean Square Error (MSE) used as a pattern matching technique, the system performance measured in terms of accuracy based on False Reject Rate (FRR) and Wrong Recognition (WR). The system performance was 100% accuracy which is high accuracy. However the system focused on recitation based on Rasm Uthmani.

Implementation of an interactive multimedia system done by [20] to learn Qur'an recitation correctly (according to Tajweed rules) and to overcome learning process problems of Qur'an recitation (limited time and limited number of teachers). The system consist of 3 levels : correction in makhraj(Hijaiyah letters), law of recitation and combination of recitation law and correction in makhraj/pronunciation. MFCC was used as feature extractor. The accuracy achieved by the system was 90%, 70% and 60% for the three levels respectively.

1. The image from [17]

Review paper of Qur'an recitation verification automatic systems was mentioned by [6]. The paper described speech recognition systems and its phases (Pre-processing, Feature Extraction, Training and Testing and Features Classification (Pattern Recognition) and structure of Qur'an recitation verification automatic systems in details (as example Verification HMM models, confidence Score Analysis and phoneme duration analysis.). Also the paper mentioned some systems that helped users in learning, memorizing and verifying Qur'an recitation such as E-Hafiz system. The paper also presented a proposal for an automated system to verify the reading of the Holy Qur'an correctly. This system will use MFCC for extracting features and HMMs for recognizing and matching features which are robust technologies in Arabic speech recognition. Another study done by [10] was a lip speech recognition system that depends on the lip movement during pronunciation of Arabic letters to recognize speech without hearing. The system compare lip movement between expert reciter and novice reciter to determine if novice reciter pronunciation was correct or not. High speed camera was used to record the data from expert reciter for 4 times (28 alphabets of Qur'an) in audio visual studio room at University of Malaysia (IIUM). Width and height of the lips were used to extract features of each frame to extract relation between them to determine the position of the mouth during pronouncing letters which categorized in four groups (Normal, Agape, Open and Stretched). Graph of lip movement and position of the mouth during recitation of the expert reciter was used as model reference of the system and the system tested by comparing between the two graphs (expert graph and novice graph) to check if the pronunciation was correct or not. Also the study presents that the pronunciations of Qur'anic letters rely on two major things : points of articulation and attributes (Sifaat) of the letters. The study explained the main places of articulations in the vocal tract : empty space in the mouth and throat, the throat, tongue, two lips and the nasal passage, there are difficulties in the pronunciation of letters that share the same point of articulations or its articulations are close to each other but these letters can be differentiated by attribute (Sifaat) of these letters.

The above diagram (fig 2) shows the articulation points and the alphabets that are related to these points.

Correctness of user recitation automatically according to Tajweed rules was a system implemented by [8] to overcome the limitation of individual learning. The system was based on a CMU Sphinx tools which is an open-source tool, due to its flexibility in creating pronunciation dictionary for Arabic letters. The acoustic model for the system built by 10 reciters for two chapters of Al- Qur'an. HMM which is robust technology was used to extract features from wave signals, these input signals were decoded in 3 parts, acoustic model part, language model part, which created using Qur'anic text, and the pronunciation Dictionary part. The system accuracy was calculated by aligning the identified words against the correct word of verse.

2. PROJECT GOAL

There are many ASR systems built and developed to facilitate the recitation and memorization of Qur'an, such as HAFSS© system which is one of the best applications handling most of the recitation rules. HAFSS© is teaching the correct recitation of the Holy Qur'an. One of its features is to detect errors in user recitation and teaching the correct recitation of the Holy Qur'an. The aim of this project is to build ASR system for Indian speakers to detect and correct errors in their recitation (pronunciation errors or Tajweed errors) to enhance the performance of HAFSS© application by using Deep Neural Networks (DNNs) based on nnet3 recipes provided by kaldi ASR toolkit, one of the best models used for speech recognition. The developed system has taken into it's consideration the non-Arabs problems with Arabic pronunciation and has been built to focus on detecting the non-proper pronunciation of uttered letters and got powerful technique to verify Qur'an recitation which gives better performance, knowing that there are some Arabic letters that non-Arabic speakers can not pronounced them properly and correctly and may be confused with other letters like (خ = كا ،ح = ها ،ع = إين ، غ = قين، ض = دا، ظ = ذا ، ط = تا ، ق = كاف) (each letter mentioned with confuced one). The developed system concentrated on these confuced letters to correct their pronunciation and after that has teachs them the recitation rules. The accuracy of error detection is one of the factors that differentiate between one ASR and another. Another factor is the speed of input manipulation. Existing applications must be improved in many ways to reach the desired accuracy and speed [9].

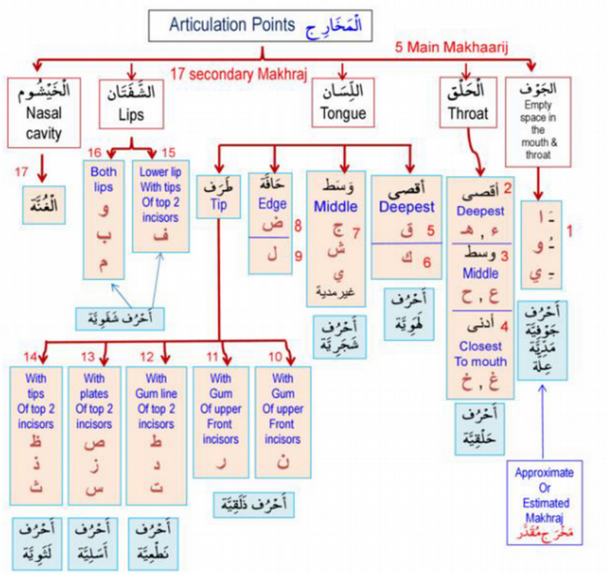

Articulation Points اَلْمَخَارِج
5 Main Makhaarij
17 secondary Makhraj
اَلْخَيْشُوم Nasal cavity
اَلشَّفَتَان Lips
اَللِّسَان Tongue
اَلْحَلْق Throat
الجَوْف Empty space in the mouth & throat
17 اَلْغُنَّة
16 Both lips و ب م
15 Lower lip With tips Of top 2 incisors ف
أَحْرُف شَفَوِيَّة
طَرَف Tip
حَافَة Edge ض 8 ل 9
وَسَط Middle ج ش ي 7 غيرمدية
أَحْرُف شَجَرِيَّة
أقصى Deepest ق 5 ك 6
أَحْرُف لَهَوِيَّة
2 أقصى Deepest ء , هـ
3 وسط Middle ع , ح
4 أدنى Closest To mouth غ , خ
أَحْرُف حَلْقِيَّة
1 ـَ ا ـُ و ـِ ي
أَحْرُف جَوْفِيَّة مَدِيَّة عِلَة
Approximate Or Estimated Makhraj مَخْرَج مُقَدَّر
14 With tips Of top 2 incisors ظ ذ ث
13 With plates Of top 2 incisors ص ز س
12 With Gum line Of top 2 incisors ط د ت
11 With Gum Of upper Front incisors ر
10 With Gum Of upper Front incisors ن
أَحْرُف لَثَوِيَّة
أَحْرُف أَسَلِيَّة
أَحْرُف نَطْعِيَّة
أَحْرُف ذَلَقِيَّة

FIG. 2 – Articulation points of each arabic letter

3. THE STRUCTURE OF THE SYSTEM

In this section the describtion of the developed system and it's steps will explained in details.

3.1 Data collection

The data, which ASR systems require, are extensive amounts of text data to train statistical language models, transcribed speech recordings from many speakers and pronunciation dictionaries, that cover the full vocabulary of the language, or at least the training corpus [21]. The data set was collected from 94 Hindi speakers

each speaker has 11 audio files, 10 files containing recording of the 10 mispronounced Arabic letters(خ ،ح ،ع ، غ ، ض ، ظ ، ط
، د ، ق ، ص) as shown in table 1, and one audio file containing recording of 5 short chapters of Qur’an (Suras) (chapters 1, 108.109,112,113 and 114), all files saved as (.wav) files. About 65 hours recording of data. 58 hours for training and 7 hours for testing at 16KHz sample rate, mono channel.
Mispronounced letters and their transcription is shown in table 1.

Latter name in Arabic	Latter name in English	Letter transcription
ضاد (ض)	Daad	/D/
ظاء (ظ)	DHA	/ Z/
تاء (ت)	TAA	/t/
طاء (ط)	TTA	/T/
حاء (ح)	HAA	/h/
خاء (خ)	KHA	/X/
عين (ع)	AIN	//
غين (غ)	GIN	/g_h/
ذال (ذ)	DHA	/z̃/
صاد (ص)	SUD	/S/

TABLE 1. – The 10 Arabic mispronounced letters.

3.2 Data preparation

The collected data set was divided in to 2 folders, training folder and testing folder, each of these 2 folders containing one folder for each speaker. The speaker folder contains his 11 recorded audio files and each audio file has it’s own transcription file which described what the speaker said in a text form. In addition of that the training and testing folders have 4 additional files as described below :

1. Utt2spk file : Consist of mapping between all utterances of speaker to speaker ID (11 utterance to one speaker)
2. Wav.scp file : Consist of absolute path of audio files (connects every utterance to its associated audio file).
3. Text file : Consists of text transcription of all utterances recorded by all speakers.
4. Spk2utt file : Consist of speaker ID mapping to his all utterances, in this study this file consist of speaker ID to 11 utterance (one speaker to 11 utterance).

The above first 3 files (Utt2spk,Wav.scp and Text file) must be created manually by yourself but Spk2utt file can be created either manually or by using special kaldi scripts that written specifically to do this. Also there is "corpus.txt" file which contains every single utterance transcription that can occur in ASR system (in this case it will be 852 lines from 852 audio files).

3.3 Features extraction

Features extraction is the step followes data preparation. Used to represent the phones within the words with neglection of other degrading factors in the signal such as the channel characteristics and nois [2]. The extraction of the best parametric representation of acoustic signals is an important task to produce a better recognition performance. The commonly used feature extraction methods is MFCCs, it is very robust feature extraction method because it based on human hearing perceptions [19]. The speech signals were sampled at 16 KHz, the feature was extracted by applying 25ms Hamming window, with a 10ms overlap (25ms frames shifted by 10ms each time) in additional delta and delta-delta coefficients. Here Mel Frequency Cepstral Coefficients (MFCCs), which are derived from FFT-based log spectra, was used. The length of the features vector was 40, the length of the parameterised static vector (MFCC0 = 13) plus the delta coefficients (+13) plus the acceleration coefficients (+13) + 1 energy coefficient [24].

3.4 Language Model (LM) creation

Language models (LMs) used to assign probabilities to sequences of words. There are some simplest LMs which assigns probabilities to sentences and sequences of words, such as n-gram order. N-gram is a sequence of N n-gram words. There are many n-gram order such as 1-gram (or unigram) is a one-word sequence of words, 2-gram (or bigram) is a two-word sequence of words and a 3-gram (or trigram) is a three-word sequence of words [16]. The popular toolkit for building LMs is SRILM toolkit . Various language modeling toolkits are used in the Kaldi example scripts, SRILM is the best one [1].
This study has used 2-gram, 3-gram and 9-gram LMs to assigns probabilities to sequences of words and then has used these LMs to build the acoustic model. The experiments results proved that the 9-gram LM was the best LM among others which gave highest accuracy and lowest WER, and also proved that the WER increase and accuracy decrease when n-gram order exceeded 9-gram.

3.5 Acoustic Model (AM) creation

An acoustic model (AM) is statistical representations of phonemes (sounds that makes up a word). The Hidden Markov Models (HMMs) is the statistical representations model that used to represent phonemes statistically. Each phoneme has its own HMM and each HMM has 3 states. In the HMM-GMM model, each state fits a frame of acoustic input. However, in this model, the input could be several frames of coefficients, and the output is HMM states based on posterior probabilities [12].
The widely used acoustic models in traditional speech recognition systems is GMM-HMM AM. In this model, GMM is used to model the distribution of the acoustic characteristics of speech and HMM is used to model the time sequence of speech signals. Since the rise of deep learning in 2006, deep neural networks (DNNs) have been applied in speech acoustic models. In 2009, Hinton and his students used feed forward fully-connected deep neural networks in speech recognition acoustic modeling [18].

3.6 Training Hidden Markov Models

To train the HMM parameters, ***A*** (the transition probabilities) and ***B*** (the observation probabilities), a training corpus, of spoken sentences, as a wave file, along with their corresponding transcriptions are needed. A pronunciation lexicon is needed to specify the phone sequences of each word in the uttered speeches [23]. Each phone is then modeled with an HMM. Given an observation sequence ***X*** and a HMM

$$\lambda = \langle \mathcal{X}, \mathcal{B} \rangle$$

the training procedure of HMMs aims to find the HMM parameters that best fit the training data.

$$\mathcal{P}\langle \mathcal{X} | \lambda \rangle$$

.

3.7 Viterbi Training

Kaldi training is based on Viterbi training algorithm, which is an approximation of the Baum-Welch training [26]. The Baum-Welch algorithm computes the probability of being in state ***i*** at time ***t***, by performing an iterative estimation to improve the HMM parameters

$$\lambda = \langle \mathcal{X}, \mathcal{B} \rangle$$

until a state of convergence is achieved. The forward and backwards probabilities for each sentence are computed in every iteration, which is the likelihood of an observation,

$$\mathcal{P} = \langle \mathcal{X} | \lambda \rangle \text{ given a HMM } \lambda = \langle \mathcal{X}, \mathcal{B} \rangle$$

and an observation sequence ***X***. To compute the observation probabilities, the forward algorithm uses a table to store intermediate values and then sums over the probabilities of every possible hidden state paths, which could have generated the observation in state ***i*** at time ***t***. The EM (Expectation Maximization) algorithm is used to train the GMM-HMM [23]. The Viterbi algorithm is less time consuming than the Baum-Welch algorithm. It chooses the most likely path of hidden states and uses it to update the hidden parameters, instead of summing over all paths that pass through a state at each time step. Because the training corpus has the information of which text transcription matches the spoken sentences, or the observations, the Viterbi algorithm can be forced to pass through specific words by setting the transition probabilities and thus, determine where in time certain words occur in the observation sequence. This is called forced alignment and the Viterbi algorithm only has to find the correct subphone sequence in order to give the best path, corresponding to the given observation sequence. The alignment of the HMM state to the observations is then used to re-estimate the HMM parameters [23]. In GMM-HMM models, the observation sequences are assumed to be generated by each hidden state according to Gaussian mixture probabilities [14].
In training the weights adjusts to find the most and best accurate path to the phoneme.
Figure 3 shows HMM states and Gussians. Each HMM state can have more than one GMM that represents the acoustic properties of each phoneme [9].

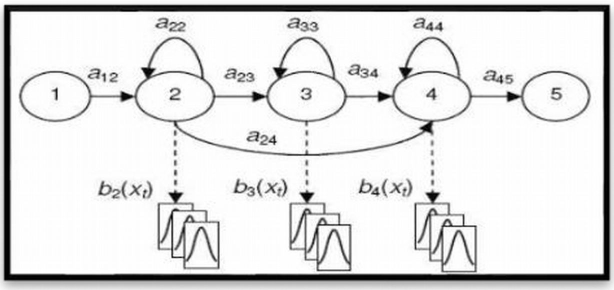


FIG. 3 – HMM states and Gussians

GMM-HMMs-based AMs are widely used in traditional ASR. Now a days DNN-HMM-based are used as AMs in ASR systems and outperforms the traditional GMM-HMM AMs, enhances the ASR performance and gives better accuracy and WERs. In the developed system the hybrid GMM-HMM and DNN-HMM were sused as AMs technique, and this gave better accuracy and improved performance but it is time and memory consuming.

3.8 Acoustic Models training

The AM training of this system was divided into 2 phases, GMM-HMM based training phase and DNN-HMM based training phase.

3.8.1 GMM-HMM -Based AMs training.

In this phase the system trained on 58 hours training set on Monophones and Triphones using 3 LMs (2-gram, 3-gram and 9-gram). Firstly the AM trained on monphone, and then the monophone AM was used to align the feature vectors of trained data. After that the monophone alignments were used to train triphone then the triphone re-aligned and a new delta and +delta-delta triphone AM, was trained and aligned. After that the LDA-MLLT (used to reduce features dimensionality and to perform de-correlation of the reduced features) was applied to the new triphone AM. Lastlly Speaker Adapted Training (SAT) was performed on top of the LDA+MLLT features and then fMLLR was applied to perform speaker normalization [21].

3.8.2 DNN-HMM-Based AMs training.

Train GMM-HMM model using training data is the first step for training DNN-HMM model. The standard Kaldi recipes for DNN-based acoustic modeling consists of the following steps (from step 1 to 5 is GMM-HMM model training) :

1. Feature extraction (13 MFCCs can be used as the features).
2. Training a monophone model training a triphone model with delta features.
3. Training a triphone model with delta and delta-delta features.
4. Training a triphone model with Linear Discriminative Analysis (LDA) and Maximum Likelihood Linear Transform (MLLT).
5. Speaker adapted training (SAT), i.e. training on feature space maximum likelihood linear regression (fMLLR) adapted features.
6. Training the final DNN-HMM model. The DNN-HMM model is trained using fMLLR-adapted features ; the decision tree and alignments are obtained from the SAT-fMLLR GMM system.

The system trained on 58 hours training set by using Time Delay Neural Networks and long Short-Term Memory neural networks (TDNN-LSTM) using nnet3 recipes provided by kaldi ASR toolkit. The TDNN-LSTM is a DNN architecture and LSTM are a special kind of Recurrent Neural Networks RNNs which have cyclic connections and also contain memory blocks in the recurrent hidden layer. The cyclic nature of the networks make them very well suited for modeling sequences, such as language modeling and handwriting recognition [22].
40-dimensional Mel-frequency cepstral coefficients (MFCCs) appended with a 100-dimensional i-vectors were used as the input of the network. Creating more training data from existing one is done by using data augmentation technique by applying speech perturbations to improve the performance of deep learning neural networks. The network was made of 8 hidden layers, which 6 of them were TDNN layers and 2 were LSTM layers, the dimensions of TDNN hidden layers and cell of a LSTM layers was set to 1024 and both

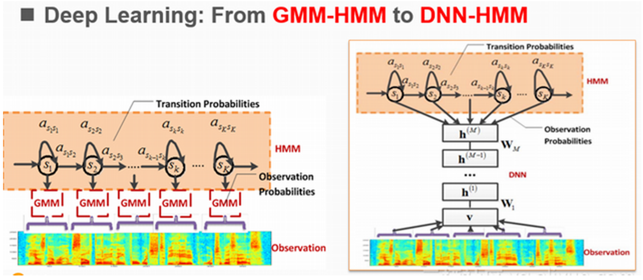


Fig. 4 – Architecture of the GMM-HMMs and DNN-HMMs.

the recurrent projection unit and non-recurrent projection unit was set to 512. ReLU (Rectified Linear Unit) function was used as an activation function. ReLU output the input directly if it is positive, otherwise, it will output zero. It has become the default activation function for many types of neural networks because a model that uses it, is easier to train and often achieves better performance [3]. Figure 4 above illustrates the architecture of the GMM-HMMs and DNN-HMMs.

3.9 Experiments Results and discussion

This section , presents the results of the system along with discussions about the results. All of the models below were tested and evaluated on 7 hours audio data set of 10 Arabic letters and 6 chapters of Qur'an recited by Indian speakers. Experimental results were reported in term of Word Error Rate (WER) which is the minimum edit distance between the output of the ASR system and the reference transcriptions (actual or correct output). The number of data set is 1004 audio files from 94 Indian speakers, devided in to 852 file for training and 152 file for testing. The models results and comparison between these results will be discussed. As mentioned before the performance measured by WER and accuracy based on kaldi standard calculations, where the WER and accuracy calculated as follow :

$$WER = 100 * \frac{\boldsymbol{I}+\boldsymbol{D}+\boldsymbol{S}}{TotalNumberofWords}$$

$$Accuracy = 100 * \frac{TotalNumberofWords - \langle \boldsymbol{I}+\boldsymbol{D}+\boldsymbol{S}\rangle}{TotalNumberofWords}$$

Where ***I*** is the number of Insertion,***D*** is the number of Deletion and ***S*** is the number of Substitution.

3.9.1 GMM-HMM AMs Results.

Table 2 presents the performance results of the GMM-HMM AMs (Monophone, Triphone, Triphone (LDA-MLLT) and Triphone (SAT) using 2-gram, 3-gram and 9-gram LMs. These results shown that the Triphone (SAT) AM using 9-gram LM outperform all other models and gives highest accuracy and low WER. Whereas figure 5 presents the accuracy of the ten Arabic letters using GMM-HMM AM.

N-Gram Order	Monophone	Triphone	Tri- MLLT	Tri-SAT
2-Gram	30.04%	20.85%	19.33%	17.43%
3-Gram	20.47%	12.70%	12.15%	10.47%
9-Gram	4.86%	4.70%	4.69%	4.68%

Table 2. – GMM-HMM AMs N-Gram Comparison

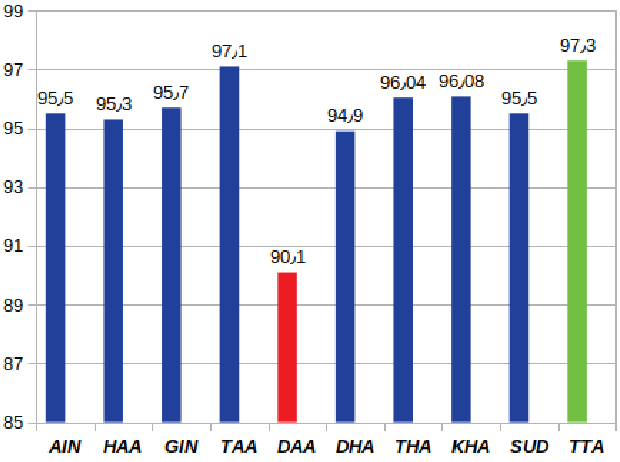


Fig. 5 – Accuracy of the ten Arabic letters using GMM-HMM

.

3.9.2 GMM-HMM Insertion and Deletion Results.

Deletion and Insertion results is another way to estimate ASR quality. Deletion occurres when the phoneme have been missed from the output word. As example one may pronounced the word (آناصعا) as (ناصع) the phoneme (") or (تنوين فتحة) was missed from the pronounced word. Whereas insertion caused when the pronunciation of word is differ from the original one due to insertion of another phoneme/s to the correct pronunciation of the word. For example one may pronounced the word (عِن) as (هِن) the letter (ع) replaced with the letter (هـ) [9].
These errors lead to phonem and word recognition errors and thus degrade the accuracy of the system.
Figure 6 shows insertion and deletion results that have been recorded per phoneme in GMM-HMM models.

3.9.3 TDNN-LSTM AMs Results.

Comparison of TDNN-LSTM AMs performance with 6 layers and 1024 hidden layers using 2-gram, 3-gram and 9-gram LMs is shown in Table 3. From the results noticed that the TDNN-LSTM AM with 6 layers and 1024 hidden layers using 9-gram LM outperformed all other models and gave highest accuracy and lowest WER, so this leads to say that the increasing of n-gram ensures more accuracy and ***n*** must not exceeded 9, as this may lead to unpredictable and adverse results.

3.9.4 DNN-HMM Insertion and Deletion Results.

As mentioned before Insertion and Deletion results can be used to estimate and measured the recognition quality. Figure 7 shows Insertion and Deletion results that have been recorded per phoneme

[1] *Tri-MLLT : training a triphone model with Linear Discriminative Analysis (LDA) and Maximum Likelihood Linear Transform (MLLT.)

[2] *Tri-SAT : (Speaker adaptive Training (SAT)) :training Tri-mllt model on Feature space Maximum Likelihood Linear Regression (fMLLR) adapted features.

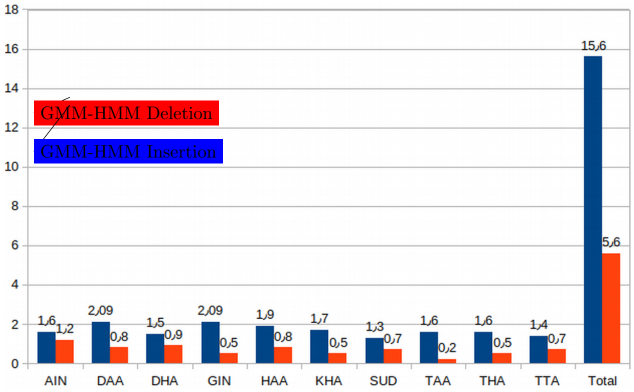


FIG. 6 – Insertion and Deletion results per phoneme using GMM-HMM model.

N-Gram Order	TDNN-HMM
2-Gram	8.9%
3-Gram	6.16%
9-Gram	3.12%

TABLE 3. – Comparsion of TDNN-LSTM AM (2-gram, 3-gram and 9-gram)

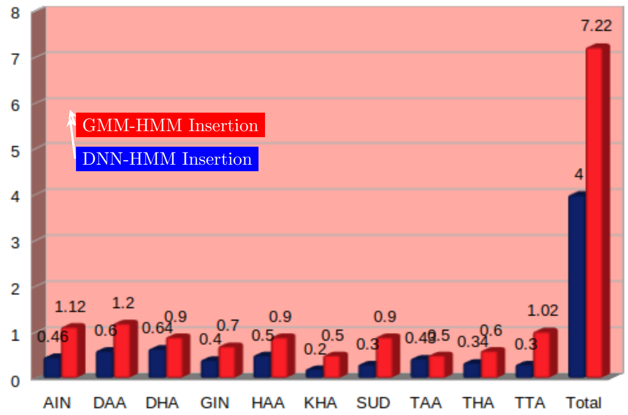


FIG. 7 – TDNN-LSTM Insertion and Deletion Results per phoneme

in TDNN-LSTM model. The figure shows that the phoneme DAA (د) has achieved highest results in both Insertion and Deletion (worest phoneme accuracy) due to confusion of this letter with the letters TAA (ت) and TTA (ط), and the phoneme KHA (خ) has achieved lowest Insertion and Deletion results (best phoneme accuracy).

4. ACOUSTIC MODELS RESULTS COMPARISON

Comparison of the performance of GMM-HMM and TDNN-LSTM AMs is shown in table 4. The results shows that the TDNN-LSTM AMs outperformed in the overall result. The TDNN-LSTM AMs were trained with different numbers of neural network layers, 3 layers with 1 LSTM layer, 6 layers with 2 LSTM layers and 9 layers with 3 LSTM layers and different hidden layers dimensions, 1024 and 2048 hidden layers. The best WER was obtained and reached by using 6 layers with 2 LSTM layers and 1024 hidden nodes per layer. The training of the GMM-HMM model was easy to learn and much faster than the TDNN-LSTM model and GMM-HMM has fewer parameters, while the training of TDNN-LSTM was memory and time consuming.

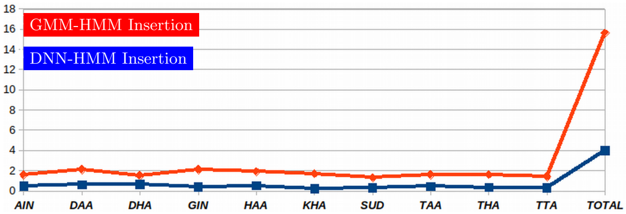


FIG. 8 – GMM-HMM and DNN-HMM Insertion results Comparison

N-Gram Order	GMM-HMM	TDNN-HMM
2-Gram	17.57%	8.9%
3-Gram	10.47%	6.16%
9-Gram	4.68%	3.12%

TABLE 4. – GMM-HMM and DNN-HMM Comparison (WER%)

Table 4 shows that TDNN–HMM outperformed GMM–HMM in all n-gram models and the best model was 9-gram LM in both GMM LM and TDNN-HMM . The difference between the two models started with 8.6% in a 2-gram and ended with 1.5% in 9-gram. Another ways to prove that DNN-HMM outperformed GMM-HMM are the Insertion and Deletion results. According to figure 8 TDNN-HMM, Insertion outperformed GMM-HMM Insertion which was 4% and 15.6% respectively.

There was a good improvement in the insertion of DAA phoneme by 2.03% when using TDNN-HMM model, also the total insertion result was improved by 11.6%. The amount of training data is small, and this is one of the reasons that led to the decrease in the accuracy of the system and the occurrence of Insertion, Deletion and Substitution problems, wherefore increasing the training data leads to solving these problems and thus increasing the system performance.

The tow figures 6 and 7 shows that deletion result of GMM-HMM outperformed DNN-HMM by 1.62% and this was unexpected result.

5. COMPARISON WITH PREVIOUSLY PUBLISHED RESULTS

There are many previously published studies in Qur'an recitation verification, these studies gave promising results. One of these studies is the system developed by [9] which used HTK toolkit to train, test and to evaluate the system. When comparing this system with developed one (such systems used the same dataset), found that the developed system outperformed the HTK system. The author investigated two models, GMM-HMM and DNN-HMM. The performance of this system was measured by computing the WER using the HTK toolkit accuracey equation.

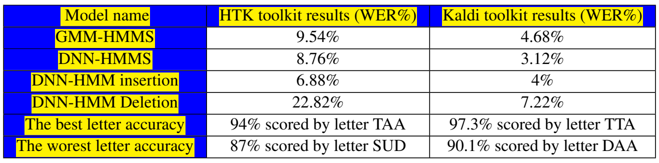
Comparing between Insertion and Deletion results per phoneme using HTK DNN and kaldi DNN, was found that kaldi toolkit beat HTK toolkit which scored 6.88% and 22.82% while kaldi toolkit scored 4% and 7.22% respectively.

When comparing the system developed by [9] and the developed one the conclution was :

* The system developed by [9] and the developed one, investigated two models, GMM-HMM and DNN-HMM.

* Both HTK tollkit model and kaldi toolkit model reached the best

accuracy 91.24% and 96.88% respectively by using 9-gram GMM-HMM.
* HTK toolkit DNN model reached the best accuracy 92.84% by using DNN-HMM architecture with three layers and 1024 hidden nodes, while by using kaldi TDNN-LSTM model the best accuracy reached was 96.88% by using TDNN-HMM architecture with six layers and 1024 hidden nodes.
* Comparing the ten letters accuracy achieved by HTK toolkit, the best accuracy scored by the letter TAA (ت) and the author found problem with the letter SUD (ص), the author mentioned that one possible reason could be the "SUD" letter is one of the so-called wheezing sounds. While by kaldi toolkit as figure 5 shows, the best accuracy achieved by the letter TTA (ط) and a problem found with the letter DAA (د), due to confusion of this letter with the letters TAA and TTA, because these 3 letters sharing the same point of articulation, this resulting in achieving this letter the highest insertion and substitution results in both GMM-HMM and DNN-HMM.
* The SUD letter accuracy using DNN-HMM model based on Kaldi toolkit outperforms the GMM-MMM model by 1% and at the same time outperforms DNN-HMM model based on HTK toolkit by 9.5%.
Table 5 illustrates and concludes the results comparison between the two systems.



Model name	HTK toolkit results (WER%)	Kaldi toolkit results (WER%)
GMM-HMMS	9.54%	4.68%
DNN-HMMS	8.76%	3.12%
DNN-HMM insertion	6.88%	4%
DNN-HMM Deletion	22.82%	7.22%
The best letter accuracy	94% scored by letter TAA	97.3% scored by letter TTA
The worest letter accuracy	87% scored by letter SUD	90.1% scored by letter DAA

TABLE 5. – Results comparison between kaldi models and HTK models systems.

6. CONCLUSION

This paper presents a hybrid GMM-HMM (Hidden Markov Model with a Mixture of Gaussians Model) and DNN-HMM (Hidden Markov Model with Deep Neural Network) system using kaldi ASR toolkit to detect and correct the pronunciation errors in Qur'an recitation for Indian speakers.
The system used about 65 hours recording of data (58 hours for training and 7 hours for testing) at 16KHz sample rate, mono channel. It focused on 10 mispronounced Arabic letters (خ ،ح ،ع ، غ ، ص ، ق ، د ، ط ، ظ ، ض).
The system investigated two acoustic models, GMM-HMM and DNN-HMM models. The performance of this system was measured by computing the WER. The performance of GMM-HMM acoustic model using 9-gram LM was 4.68% while by using TDNN-LSTM acoustic model with 6 layers, 2 LSTM and 1024 hidden nodes per layer, it was 3.12%. From the expriements results noticed that, the DNN-HMM outperforms GMM-HMM based on kaldi toolkit by 1.56%. The developed system enhancing the performance of pronunciation error detection and correction according to system results.
Compairing the developed model with HTK-HMM based model, noting that the two models used the same dataset, found that the developed one outperforms the HTK system by 5.64%. Comparison between Insertion and Deletion results per phoneme for the tow systems showed that kaldi toolkit outperformed HTK toolkit which scored 6.88% for Insertion and 22.82% for Deletion while kaldi toolkit scored 4% for Insertion and 7.22% for Deletion. The experiments results proved that the DNN-HMM using kaldi toolkit outperformed DNN-HMM using HTK toolkit. And using DNN-HMM in HAFSS© instead of GMM-HMM has enhance the performance of it. So be usefull to using DNN-HMM in ASR systems because this enhancing detection and correction of pruounciation errors in Qur'an recitatin and will give promising results.

7. RÉFÉRENCES

[1] Creating the language model or grammar. https://kaldi-asr.org/doc/data_prep.html. Accessed : 2020-11-22.

[2] The dummy's guide to mfcc. https://medium.com/prathena/the-dummys-guide-to-mfcc-aceab2450fd. Accessed : 2020-09-25.

[3] A gentle introduction to the rectified linear unit (relu). 2019. https://machinelearningmastery.com/rectifiedlinear/activationfunctionfordeeplearningneural-networks. Accessed : 08/20/2020.

[4] Mohamed S Abdo, AH Kandil, AM El-Bialy, and Sahar A Fawzy. Automatic detection for some common pronunciation mistakes applied to chosen quran sounds. In *2010 5th Cairo International Biomedical Engineering Conference*, pages 219–222. IEEE, 2010.

[5] Sherif Mahdy Abdou, Salah Eldeen Hamid, Mohsen Rashwan, Abdurrahman Samir, Ossama Abdel-Hamid, Mostafa Shahin, and Waleed Nazih. Computer aided pronunciation learning system using speech recognition techniques. In *Ninth International Conference on Spoken Language Processing*, 2006.

[6] Ayat Hafzalla Ahmed and Sherif Mahdi Abdo. Verification system for quran recitation recordings. *International Journal of Computer Applications*, 163(4) :6–11, 2017.

[7] Alaa N Akkila and Samy S Abu-Naser. Rules of tajweed the holy quran intelligent tutoring system. 2018.

[8] Ahmed AbdulQader Al-Bakeri and Abdullah Ahmad Basuhail. Asr for tajweed rules : Integrated with self-learning environments. *International Journal of Information Engineering & Electronic Business*, 9(6), 2017.

[9] Mubarak Al-Marri, Hazem Raafat, Mustafa Abdallah, Sherif Abdou, and Mohsen Rashwan. Computer aided qur'an pronunciation using dnn. *Journal of Intelligent & Fuzzy Systems*, 34(5) :3257–3271, 2018.

[10] Tareq Altalmas, Muhammad Ammar Jamil, Salmiah Ahmad, Wahju Sediono, Momoh Jimoh E Salami, Surur Shahbudin Hassan, and Abdul Halim Embong. Lips tracking identification of a correct quranic letters pronunciation for tajweed teaching and learning. *IIUM Engineering Journal*, 18(1) :177–191, 2017.

[11] S Hamid and Mohsen Rashwan. Automatic generation of hypotheses for automatic diagnosis of pronunciation errors. In *Proceedings of NEMLAR International Conference on Arabic Language Resources and Tools*, pages 135–139, 2004.

[12] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdelrahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent

Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition : The shared views of four research groups. *IEEE Signal processing magazine*, 29(6) :82–97, 2012.

[13] Noor Jamaliah Ibrahim, Mohd Yamani Idna Idris, Zaidi Razak, and Noor Naemah Abdul Rahman. Automated tajweed checking rules engine for quranic learning. *Multicultural Education & Technology Journal*, 2013.

[14] Noor Jamaliah Ibrahim, Mohd Yamani Idna Idris, and Zulkifli Mohd Yusoff. Computer aided pronunciation learning for al-jabari method : A review. *QURANICA-International Journal of Quranic Research*, 6(2) :51–68, 2014.

[15] Ahsiah Ismail, Mohd Yamani Idna Idris, Noorzaily Mohamed Noor, Zaidi Razak, and Zulkifli Mohd Yusoff. Mfcc-vq approach for qalqalahtajweed rule checking. *Malaysian Journal of Computer Science*, 27(4) :275–293, 2014.

[16] Dan Jurafsky. *Speech & language processing*. Pearson Education India, 2000.

[17] SEHM Metwalli. Computer aided pronunciation learning system using statistical based automatic speech recognition techniques. *Faculty of Engineering at Cairo University in Partial Fulfillment of the Requirements for the Degree of DOCTOR OF PHILOSOPHY in ELECTRONICS AND COMMUNICATION ENGINEERING FACULTY OF ENGINEERING, CAIRO UNIVERSITY GIZA*, 2005.

[18] Abdel-rahman Mohamed, George Dahl, and Geoffrey Hinton. Deep belief networks for phone recognition. In *Nips workshop on deep learning for speech recognition and related applications*, volume 1, page 39. Vancouver, Canada, 2009.

[19] Lindasalwa Muda, Mumtaj Begam, and Irraivan Elamvazuthi. Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques. *arXiv preprint arXiv :1003.4083*, 2010.

[20] Budiman Putra, B Atmaja, and D Prananto. Developing speech recognition system for quranic verse recitation learning software. *IJID (International Journal on Informatics for Development)*, 1(2) :14–21, 2012.

[21] Anna Vigdís Rúnarsdóttir. *Re-scoring word lattices from automatic speech recognition system based on manual error corrections*. PhD thesis, 2018.

[22] Haşim Sak, Andrew Senior, and Françoise Beaufays. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. *arXiv preprint arXiv :1402.1128*, 2014.

[23] MA Sherif, A Samir, AH Khalil, and R Mohsen. Enhancing usability of capl system for quran recitation learning. INTERSPEECH, 2007.

[24] Thomas Hain Steve Young, M.J.F. Gales and Xunying Liu. *The HTK Book (version 3.5a)*. Cambridge University Engineering Department, 2015.

[25] R Stuckless. Developments in real-time speech-to-text communication for people with impaired hearing. *Communication access for people with hearing loss*, pages 197–226, 1994.

[26] Hassan Tabbal, W El Falou, and B Monla. Analysis and implementation of a" quranic" verses delimitation system in audio files using speech recognition techniques. In *2006 2nd International Conference on Information & Communication Technologies*, volume 2, pages 2979–2984. IEEE, 2006.

[27] AN Wahidah, MS Suriazalmi, ML Niza, H Rosyati, N Faradila, A Hasan, AK Rohana, and ZN Farizan. Makhraj recognition using speech processing. In *2012 7th International Conference on Computing and Convergence Technology (ICCCT)*, pages 689–693. IEEE, 2012.