Improving Holy Qur'an recitation system using Hybrid Deep Neural Network-Hidden Markov Model approach

Mustafa Abdallah[1, a], Mubarak Al-Marri [2,b], Sherif Abdou[3,c] ,Hazem Raafat[2,d] Mohsen Rashwan[1,e], Mohamed A. El-Gamal[1,f]
[1]Faculty of Engineering, Cairo University, Egypt
[2]Computer Science Department, Kuwait University, Kuwait
[3]Faculty of Computer Science, Cairo University, Egypt
[a]m.a.elhosiny@eng.cu.edu.eg, [b]aljunobi@hotmail.com, [c]sabdou@rdi-eg.com, [d]hazem@cs.ku.edu.kw [e]mrashwan@rdi-eg.com , [f]mhgamal@aucegypt.edu

ABSTRACT

Teaching Holy Qur'an recitation rules and Arabic pronunciations to non-native speakers is a challenging task. Automatic Speech Recognition (ASR) utilizing Machine Learning techniques proved to be very promising. In this paper, we carried out a large number of experiments to achieve a significant improvement in the accuracy of an ASR system. A hybrid Deep Neural Network-Hidden Markov Models (DNN-HMM) approach is used for that purpose. Comparing the Recognition performance of the proposed approach with the traditional baseline HMM approach is performed. It turns out that our proposed approach is superior considering phone Error rate (PER). Experimental results show a significant improvement of the proposed approach in terms of recognition performance. Moreover, the performance of rules like (Vibration, Assimilation, Turning, etc.) is also improved. The proposed approach is tested using N-gram Language Model and Lattice Network.

***Keywords*: Hidden Markov Model, Automatic Speech Recognition, Deep Neural Network.**

1. INTRODUCTION

Using machine learning in developing ASR for the sake of helping non-native Muslims to learn Holy Qur'an recitation rules is widely used in recent years. In particular, a commercial system for automatic assessment of recitation of the Holy Qur'an "HAFSS©" was developed in (Sherif Abdou et al., 2006). This system implemented a speech-enabled Computer Aided Pronunciation Learning (CAPL) system. However, upgrading and improving the system performance was very much needed. An enhancement of the system was proposed in (Abdurrahman Samir et al., 2007). In this research, a modification of the baseline system was suggested to reduce the amount of the enrollment time while keeping the system accuracy at the same level. The correlation between the judgments of HAFSS system and the judgments of human experts was also measured.

In this paper, a hybrid DNN-HMM technique to build a robust ASR system to improve Qur'an recitation's performance is presented. Most of the factors that affect the recognition performance in Quran's recitation task are analyzed and studied. In section 2, Traditional HMM approach is presented. In section 3, the proposed DNN-HMM technique is fully

described. Section 4 shows different results and analysis of its impact. Comparison between the used approaches will be discussed in section 5. Conclusions and future work will be discussed in details in section 6.

2. TRADITIONAL HMM APPROACH

A significant research work was carried in Speech Recognition and Speech Verification systems using HMM models. A large number of papers published discuss using different HMM techniques for many tasks such as large vocabulary tasks and Speech Engines. Many techniques were used to train HMM models for phonetic classification. Large-margin Hidden Markov Model work was proposed in (F. Sha et al., 2006). HMM is used in Arabic-based speech verification system in (Omar et al., 1999). In this work the Arabic phoneme set was clustered into a group of clusters and the pronunciation assessment was accomplished.

As a matter of fact, some ASR tools are available to be used in building an ASR system for any language. The Hidden Markov Model Toolkit (HTK) is used in building the HMM baseline model implemented in our paper.

2.1 Dataset Description

Data was collected from 100 speakers in India, and they were divided into 80 speakers for training and 15 for testing (the remaining 5 utterances was labeled as noisy).Each Speaker records eleven utterances, so the overall utterances used in training is about 830. The dataset was manually transcribed by experts. The duration of training data is about 35 hours of speech which includes single phoneme, words and verses from the Holy Quran. Letters that might cause problem to the non-Arab Muslims are emphasized. Each utterance contains 66 examples for the letter so all recitation rules are included in our dataset. These Arabic letters with their symbols are listed in Table 1.

Table 1: Phonetic Symbols of Letters

Letter in Arabic	Symbol in English	Phonetic Symbol
الضاد	Daad	/~d/
الظاء	DHA2	/~Z/
التاء	TAA	/ t /
الطاء	TTA	/ T/
الحاء	HAA	/~h/
الخاء	KHA	/x /
العين	AIN	/~@/
الغين	GIN	/g_h/
الذال	ZAA'	/~z/
الصاد	SAAD	/S/

2.2 Experimental Setup

Extracted features – in most of our experiments – from speech utterances were Mel Frequency Cepstral Coefficients (MFCCs) and energy, along with their first and second temporal derivatives. As such, the length of the feature vector is of dimension 42. Extracted features are then normalized to make the whole training data as a real Gaussian random

variable with zero mean and unit variance. On the other hand, Filter-Bank Coefficients (FBANK) and energy are also extracted in other experiments with their first and second temporal derivatives. Therefore, the length of the feature vector is of dimension 123. These features are also normalized to have the zero mean and unit variance as well.

Training was done using five parallel clients and a main server on a core-i7(TM i7-2670QM) Dell machine that was faster than serial implementation. Re-Estimation of the model parameters was done by partitioning the data to chunks and evaluate them before collecting all chunks to a global file for final adjustment of the parameters.

3. HYBRID DNN-HMM APPROACH

Neural Networks have been widely used in speech modeling in recent years. In this context, (Abdel-rahman Mohamed et al., 2012) proposed Acoustic Modeling using Deep Belief Networks. The hybrid DNN-HMM approach in Acoustic Modeling was also introduced in (George E. Dahl et al., 2012). In this research, a new technique which depends on using Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition was proposed. Moreover, (Ossama Abdel-Hamid et al., 2012) investigated the application of convolutional neural network to build a hybrid DNN-HMM model. The previous approaches have been implemented successfully in IBM, Microsoft and Google Research labs on thousands of hours of acoustic speech.

In this paper, we propose a hybrid DNN-HMM to improve the performance of the Qur'an Recitation system. This proposal marks the first application of such approach in Arabic datasets. A detailed schematic representation of the proposed approach is shown in Figure. 1.

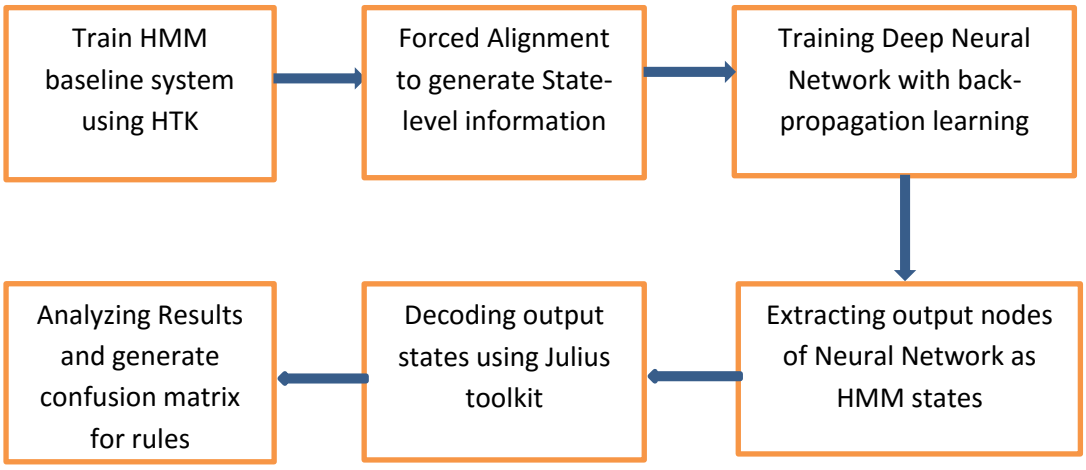

Figure. 1 DNN-HMM proposed model for Holy Qur'an Recitation system

3.1 Experimental Setup

First, extracted features MFCCs after normalization that used in creating feature vector of dimension 42 as indicated in previous section. The label of each frame is then extracted from the state-level alignment of the baseline HMM built model. In our setup, output layer has 186 target class labels (i.e., three states for each one of the 62 phones) including all Arabic

phonemes and Recitation rules. Finally, input was represented as context frames. A large number of experiments were performed in which the number of hidden layers, the number of neurons for each hidden layer and the number of context frames are changed.

3.2 Computational Setup

Training DNN on CPU is computationally expensive. As such, GPU was utilized to make use of Parallel Processing capabilities using Cuda, which is a programming language. Cuda can access Hardware and divides processors to Grids. Training was accelerated by using graphics processor unit (Tesla-C2075 GPU) via the CUDAMAT library (V. Mnih 2009). A single Iteration ("epoch") for a neural network with four hidden layers and 1096 units per layer took about 15 minutes. The single GPU is about 25 times faster than a single 2.20 GHz Intel(R) core-i7 implementation was taking about 5 hours to finish a single epoch. Decoding DNN-HMM was done by two ways, namely Julius toolkit implemented in (Lee, Kawahara et al., 2001) and Matlab code.

4. RESULTS AND DISCUSSION

For training and testing HMM, there are four factors that have an effect on the recognition performance of the acoustic model:

1. Features of the input wave files
2. Word insertion penalty and language model scale factor
3. Number of Gaussian mixtures per state
4. Number of states per phoneme

Two main features were used for training that are: MFCC and FBANK input features as mentioned above. For a word insertion penalty at thirty and language model scale factor at five. The testing results of HMM model are listed in Table 2.

Table 2: Comparison of different features for HMM

Features	MFCC	FBANK
Classification Accuracy	90.46 %	83.98 %

As far as the number of Gaussian mixtures per state is concerned, several models were built starting from four Gaussian mixtures per state until thirty-two mixtures per state. It turns out that 30 mixtures gave the best performance as shown clearly in Figure. 2. Finally, three states per phone gave the best performance.

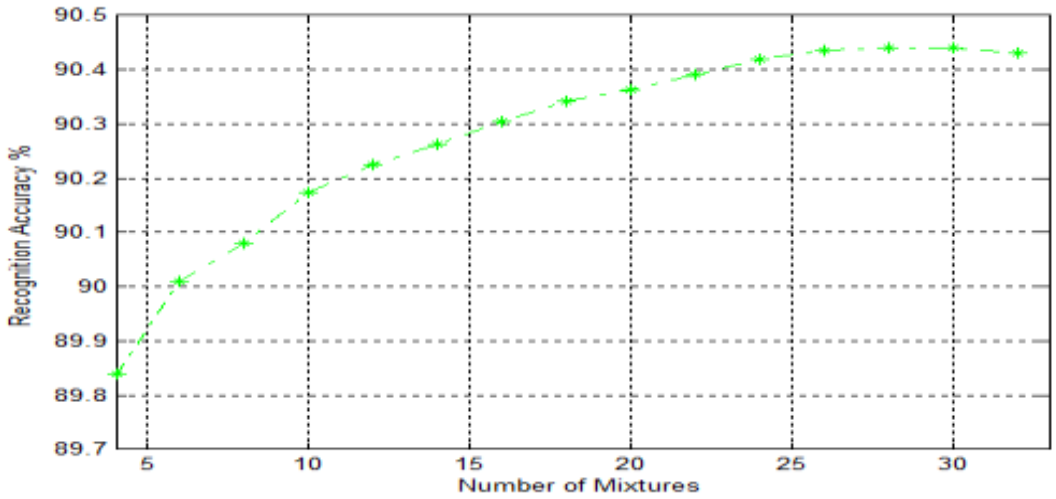

Figure. 2 Effect of increasing Gaussian mixtures on phoneme accuracy level

On the other hand, for training and testing Hybrid DNN-HMM system, there are three major parameters that affect the performance of the hybrid system:

1. Number of context frames in input
2. Parameters of decoding of posterior probabilities
3. Number of neurons in each hidden layer and number of hidden layers

In this research, the number of context frames can significantly increase the performance of the system. Many neural networks were trained starting from a single frame until 51 context frames. This parameter increases the accuracy by more than 3 percent in the case of bigram language model with three hidden layers and 1024 nodes which is shown clearly in Figure. 3.

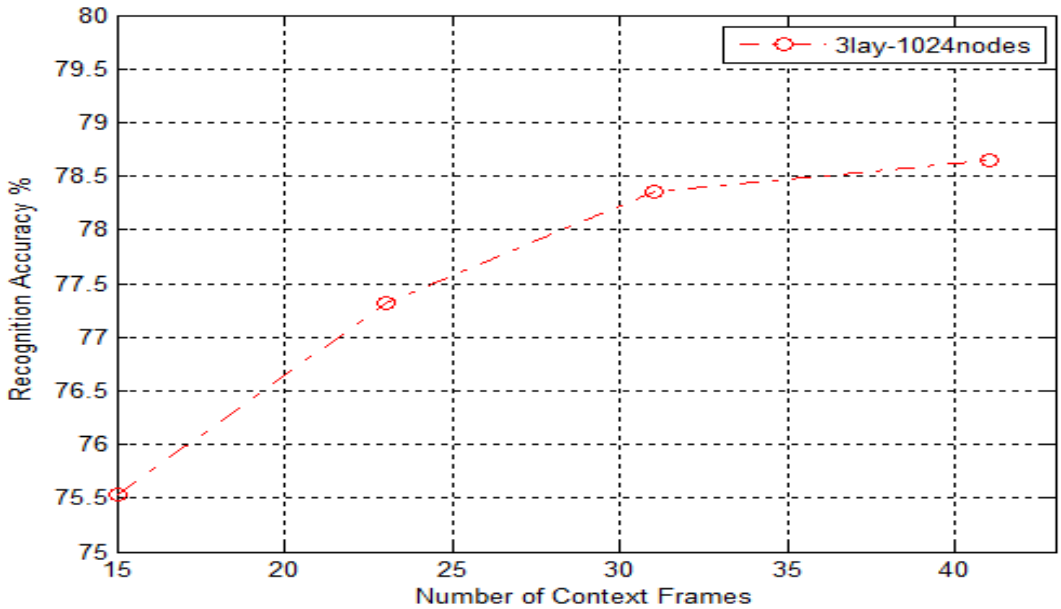

Figure 3 Effect of increasing Context Frames on phoneme accuracy level

On Decoding using Julius toolkit, the best configuration was achieved with a penalty of -50 of the first pass and -30 to the second pass. If second pass doesn't converge, result of first pass is taken which is called "fallback1pass. "

Decoding time is also an important factor for the decoding parameters that was slightly larger for the hybrid system than the baseline system decoding time comparison is made in Table 3.

Table 3: Comparison in the aspect of decoding time

Toolkit used	Proposed approach	Decoding time for one utter.
Julius	HMM	43 seconds
HTK	HMM	35 seconds
Julius	DNN-HMM	85 seconds

The number of neurons per hidden layer is a main factor in this work. A lot of experiments considering changing number of neurons per hidden layer was performed. Increasing number of nodes from 128 to 4k nodes were done which is shown in Figure. 4.

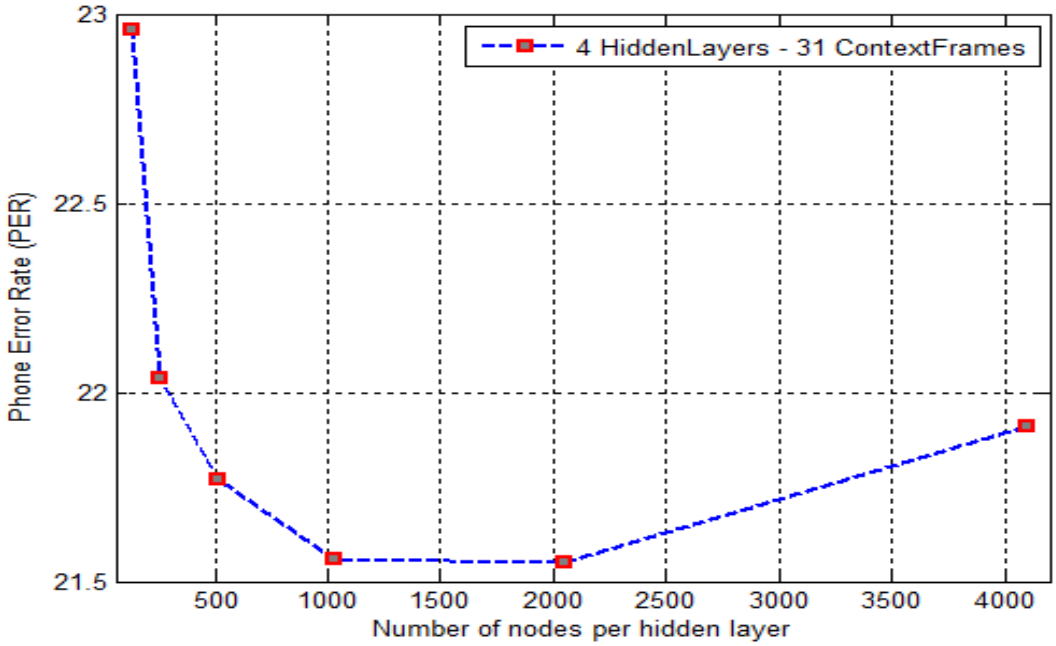

Figure. 4 Effect of increasing number of nodes per layer on phoneme accuracy level

5. COMPARISON OF THE PROPOSED APPROACH AND THE BASELINE HMM APPROACH

For the sake of comparison, an N-gram language model method is used. An N-gram is a probabilistic language modeling based on counting the occurrences of each phone in the training corpus and giving a probability to each possible transition between phonemes. In particular, Bi-gram and tri-gram language models. Julius toolkit was used for that method since it supports until ninth-gram Language Model. DNN-HMM gives about 23% relative increase over HMM in case of bigram. Increasing the order of language model decreases the difference between the two systems as shown in Table 4.

Table 4: N-gram comparison in terms of PER

N-gram	DNN-HMM	HMM system
2-gram(Bi-gram)	21.4	27.85
3-gram(trigram)	18.07	22.55
9-gram	8.76	9.6

Moreover, the lattice of each word is used in the comparison. Lattice contains all possible recitation errors and pronunciations errors of the word. The decoder output depends on the input lattice of the example and the acoustic model. The overall accuracy of all words in our data are given in Table 5. It is apparent that DNN-HMM gives about 14% relative increase over HMM in case of using Lattice networks in decoding.

Table 5: Lattice comparison in terms of PER

Used Approach	DNN-HMM	HMM system
Lattice decoding results	7.16	8.18

Finally, measuring the performance of Recitation's rules is a practical measure of the performance of the proposed system. This was extracted from the confusion matrix generated from recognition output. Samples of performance of the rules are listed in Table 6.

Table 6: Performance of some recitation rules

Recitation Rule Abbrev.	Phonetic symbol	Hybrid DNN-HMM
Pronounced N.	/n/	86.4 %
Assimilation	/n_1/	81.3 %
Pronounced M.	/m/	85.9 %
Vibration	/k_l/	72.5 %
Y- stretch	/i_2/	84.2 %
W- stretch	/u_2/	88.1 %

6. CONCLUSION

The work presented in this paper marks the first usage of DNN-HMM hybrid approach in the Holy Qur'an recitation system. Improving the performance of the Acoustic model utilizing this approach via two levels of testing, namely, the language model and Lattice's decoding methods. An NVidia Graphics processor was used to speed up the training of Deep Neural Network. It showed a significant improvement in the training time of the hybrid system. A comparative study was also done on the baseline HMM system and the proposed DNN-HMM system to determine the best configuration for our task. Decoding time for the hybrid system was slightly larger than the baseline system that is going to be addressed in our future work.

7. ACKNOWLEDGMENTS

Special thanks are posed to The Research & Development International Company (RDI®) for its great support of this work. Authors appreciate Dr. Abd El-rahman Samir for his efforts and helping us with cudamat implementation of back propagation algorithm and good

technical advices. Also we would like to thank Saeed Abdo, Mr. Khaled and all speech technology and linguistic support teams at RDI® for their valuable efforts in transcribing and revising data which was very important task.

8. REFERENCES

S. Abdou, S. Hamid and M. Rashwan, (2006). Computer Aided Pronunciation Learning System Using Speech Recognition Techniques, INTERSPEECH 2006, pp.849–852.

Abdurrahman Samir and Sherif Mahdy Abdou (2007). Enhancing usability of CAPL System for Qur'an recitation learning, INTERSPEECH 2007, pp.214–217.

F.Sha and L.Saul (2006).Large margin Gaussian mixture modeling for phone classification andrecognition, ICASSP, 2006, pp.265–268

Omar, M. K (1999).Phonetic segmentation of Arabic speech for verification using HMM, M.Sc. thesis, Cairo University, Faculty of engineering, Department of Electronics, Egypt, 1999.

Abdel-rahman Mohamed , George E. Dahl, and Geoffrey Hinton (2012). Acoustic modeling using Deep Belief Networks, IEEE Trans. on Audio, Speech, and Language Processing Vol. 20, NO. 1, pp.1-10.

George E. Dahl and Dong Yu (2012). Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition , IEEE Trans. On Audio, Speech, and Language Processing, VOL. 20, NO. 1, pp.30–42.

Ossama Abdel-Hamid, Abdel-rahman Mohamed and Hui Jiang (2012) Applying Convolutional Neural Networks concepts to hybrid NN-HMM model for speech recognition, ICASSP 2012, pp. 4277-4280.

V. Mnih (2009). Cudamat : a CUDA-based matrix class for python, Department of Computer Science, University of Toronto, Tech. Rep. UTML TR 2009-004, November 2009.

Lee, Kawahara, et al. (2001). Julius – An open source real-time large vocabulary recognition engine. In Proc. European Conf. Speech Comm. & Tech. (EU-ROSPEECH), pp.1691-1694.