Asian Journal of Information Technology 18 (2): 49-56, 2019
ISSN: 1682-3915

Large Vocabulary Arabic Continuous Speech Recognition using Tied States Acoustic Models

Mona A. Azim, A. Aziz A. Hamid, Nagwa L. Badr and M.F. Tolba
Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt
monaabdelazim@cis.asu.edu.eg

Abstract: The Hidden Markov Model (HMM) lies at the heart of the modern speech recognition systems as it provides a simple, effective and straight forward frame work to model the time varying acoustic features of the speech signals. The basic process of building HMM based speech recognition systems is a straight forward process. Nevertheless, the proper parameter estimation of such models requires large training data. Therefore, parameter tying techniques were developed to reduce the parameters of HMMs without affecting the overall system performance. This study proposes an Arabic phonetic decision tree necessary to build Tied State tri-phone HMMs. Experimental results show promising word correctness when compared with both data driven tri-phone models and phoneme based models. The maximum word correctness achieved by the proposed approach was 95.13%. Whereas it reached 78.03 and 58.45% using data driven tri-phones and phoneme based HMMs, respectively, when tested on the same benchmark database.

Key words: Straight forward, speech recognition, tri-phone, speech signals, benchmark, Arabic phonetic

INTRODUCTION

Arabic language is the most spoken language within the Semitic languages and is spoken by more than 422 million native speakers located mainly in the Arab world. Also, it is one of the six languages approved within the United Nations. In terms of the internet usage, Arabic language has occupied the fourth place in the internet usage estimates by language as it is being used by more than 168 million users. Arabic language is a very rich language in terms of vocabulary and linguistic roots as it contains more than 16,000 roots while English and Hebrew contain about 4,000 and 2,500 roots, respectively.

The technological revolution that we are experiencing now a days has encouraged the researchers to develop new technologies to facilitate the interaction between users and machines. Whereas speech is the primary means of communication between people it is preferable to be used to communicate with computers as well. That is why the development of Arabic speech recognition systems has flourished recently resulting in promising outcomes in this field. Other usage of speech recognition systems can be found in various fields such as home automation, pronunciation evaluation in computer aided language learning systems, robotics, automatic translation and sub-titling.

Most of the latest Large Vocabulary Continuous Speech Recognition (LVCSR) systems use context dependent Hidden Markov Models (HMMs) to model speech data. In order to model the differences in speaker characteristics and pronunciations, it is common for the recognition system to include several parameters which need to be estimated using huge amounts of speech training data. This increase in the number of parameters is due to the need to model acoustic units in terms of their context. However, many acoustic contexts are not observed with appropriate frequency in the training data and therefore, estimating model parameters for each context dependent acoustic unit is difficult (Young, 1992).

With continuous density HMMs which are commonly used in state-of-the-art LVCSR systems, phonetic state tying is used to reduce the total number of HMM parameters needed to be estimated while retaining the model accuracy. These can be implemented with traditional data driven techniques such as k-means clustering or knowledge driven techniques such as phonetic decision trees. Decision tree based state tying has recently gained popularity due to its successful application to several speech recognition tasks with a wide range of complexity (Odell, 1995, Beulen and Ney, 1998; Lazarides *et al*., 1996; Reichl and Chou, 1998). The decision tree based state tying algorithm uses both the training data as well as phonetically derived questions to cluster the models states. It is also capable of estimating HMM parameters of the contexts that rarely occur in the training data.

Corresponding Author: Mona A. Azim, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt, monaabdelazim@cis.asu.edu.eg

In this study, we propose a Large Vocabulary Arabic Continuous Speech Recognition System (LVCSR). The developed system used context dependent HMMs to model the acoustic features of speech signals. To improve the system performance, state tying techniques were used to tie the models states. Both state tying techniques were applied on the built acoustic models. A newly developed Arabic phonetic decision tree was built and used in the tree based tying state experiment. In order to measure the effect of using the newly created Arabic phonetic decision tree the performance of the system was observed using both types of tying techniques.

Literature review: The research of Arabic continuous speech recognition systems is a multi-discipline research field that requires integrating different aspects of research fields such as Arabic phonetics. Speech processing techniques Elshafei (1991) and natural language (Farghaly and Shaalan, 2009; Elshafei *et al*., 2002). Developing Arabic continuous speech recognition systems has gained many of the researcher's interest recently.

By Fahad and Otaibi (2001) Abushariah *et al*. have introduced an efficient and effective framework for developing speaker independent Arabic continuous speech recognition system. The developed system used the text corpus produced by KACST to generate the speech corpus by recording the 367 sentences in the KACST text corpus. The system was developed using CMU sphinx tool. Both the phoneme based and tri-phone based models were developed. The tri-phone models were generated and the states with similar distribution are tied together. The system obtained a word recognition accuracy of 92.67 and 93.88% for similar speakers with different sentences, respectively.

Arabic continuous speech recognition tasks usually address recognizing Arabic digits or Arabic alphabet broadcast news transcriptions. They explored several state-of-the-art systems and tools for Arabic speech recognition. Arabic digits recognition systems were implemented by Alotaibi (2008), Hyassat and Zitar (2006) and Satori *et al*. (2007). The system by Hyassat and Zitar (2006) used CMU sphinx engine based on HMMs which obtained a word recognition rate of 99.21% for about 35 min of training speech data and 7 min of testing speech data. The system by Satori *et al*. (2007) used the same engine based on HMM for the same task and obtained a word recognition rate of 85.56% for male speakers and 83.34% for female speakers. By Alotaibi (2008) a different kind of speech data is presented for recognizing Arabic digits using telephone Saudi accented Arabic corpus. The system used cambridge HTK tools based on HMMs and scored a correct digit recognition rate of 93.67%.

In addition, the Holy Qur'an was also considered for Arabic speech recognition by Hyassat and Zitar (2006). The system developed by Hyassat and Zitar (2006) used Sphinx-IV engine based on HMMs and reported a word recognition rate of 70.81% and a WER of 40.18% for corpus of 18.35 h. However, the system by Mourtaga *et al*. (2007) used HTK tools to build HMMs and achieved an average word recognition rate of 78.6%. On the other hand, the Arabic speech recognition system using broadcast news corpus was developed by Elhadj *et al*. (2014). The system was trained using about 7 h of speech using Sphinx 3 tools based on HMMs and tested using half an hour of speech. The system obtained a correct word recognition rate of 90.78, 93.04% and a WER of 10.87, 8.61% with and without diacritical marks. Other Arabic automatic speech recognition systems were developed for diverse tasks such as by Hyassat and Zitar (2006), Azmi and Tolba (2008), Choubassi *et al*. (2003) and Nofal *et al*. (2004).

A command and control system for 30 words was developed by Hyassat and Zitar (2006) using Sphinx-IV engine based on HMMs and obtained a word recognition rate of 98.18% whereas an Arabic speech recognition system using Recurrent Neural Networks (RNN) (Choubassi *et al*., 2003) was developed for recognizing six isolated words obtaining an average recognition rate of 95.58% and an Arabic ASR system to recognize 16 Egyptian proverbs was developed (Azmi and Tolba, 2008) based on HMMs using HTK and obtained word recognition rates of 56.8, 66.65 and 81.79% using monophonic, tri-phonic and syllable based recognition, respectively.

In addition, an acoustic training system for speaker independent continuous Arabic speech recognition system based on HMMs and different language models (bigram and context free grammar) using HTK was developed by Nofal *et al*. (2004). The system was implemented to investigate the WER as a function of vocabulary size for different types of language models. For bigram based language model, the system achieved a WER of 5.26 and 2.72% for 1340 and 306 words, respectively while for context free grammar, the system achieved a WER of 0.19 and 0.99% for the same words sets.

State tying techniques were introduced to improve the models performance. It refers to representing the states that share the same set of parameters with only single states by tying them together. This also can be defined as state clustering as we need to find which states are similar enough to be tied. The most common techniques used in state clustering include data driven and phonetic decision tree clustering algorithms.

The data driven state clustering approach relies on measuring the distance among all the states of HMMs

then combining them into clusters based on the distance matrix obtained. Although, the data driven based approaches are suitable for multi-lingual speech recognition systems they are not capable of dealing with the unseen contexts (Zgank *et al*., 2005; Nahar *et al*., 2013; Imperl *et al*., 2003; Nahar *et al*., 2015).

The phonetic tree based state tying was introduced by Odell *et al*. (1994) in which the developed recognition system proved that using a phonetic decision tree in state tying was as effective as the data driven technique and has a key advantage of providing a mapping for unseen tri-phones.

MATERIALS AND METHODS

LVCSR system: The main goal of the speech recognition system is to recognize the uttered word sequence $W = w_1, w_2, ..., w_N$ from the observed speech signals. In other words, we need to estimate $Y = \arg\max_w P(W|X)$ where, X is the set of acoustic features extracted from the observed speech signal. Using the Bayes rule this problem can be formulated to:

$$y = \arg\max_w (W|X) = \arg\max_w p(X|W) * p(W) \quad (1)$$

where, the probabilities p(X|W) and p(W) can be obtained from the acoustic models and language model, respectively. Figure 1 shows the typical architecture of LVCSR system as well as its main components. The details of each component are described below.

Acoustic modeling: The type of the HMM Model is defined by the base unit used in building such as word based, syllable based or phoneme based models, then the recognizer uses the built models as a part of the decoding process to guess the speech uttered along with the acoustic features of the input speech signals. Building HMM Models face multiple challenges as it needs sufficient training data to accurately estimate the models parameters and that is why state tying techniques were developed to reduce the number of model parameters that should be estimated.

Language modeling: Language Models (LM) are statistical models that assign probabilities to a sequence of words. The LM models are used in speech recognition systems to distinguish between words and phrases that are acoustically similar and to resolve the ambiguities that may result from the acoustic models (HMMs) and pronunciation models (phonetic dictionary). The N-gram language model refers to LM that is used to estimate the probability of the last word in the N-words sequence given previous N-1 words. Given a sequence of recognized words ($w_1, w_2, ..., w_n$) using the Chain rule of probability the $P(w_1, w_2, ..., w_n)$ can be calculated using Eq. 2:

$$P(w_1, w_2, ..., w_n) = P(w_1).P(w_2|w_1), ...,$$
$$P(w_n|w_1, ..., w_{n-1}), = \Pi_{k=1}^{n} P(w_k|w_1^{k-1}) \quad (2)$$

Phonetic dictionary: Phonetic dictionary provides a mapping between Arabic words and its corresponding list of phonemes. The phonetic dictionary is used by the recognizer to compose a sequence of meaningful words from the recognized phonemes resulting from testing the acoustic models. The phonetic dictionary is an essential part in the recognition system, so, it is preferable to be built manually to guarantee its high quality instead of applying a pre-defined set of phonetic rules on the text corpus.

Feature extraction: The purpose of the feature extraction process is to map the audio signals to a set of acoustic

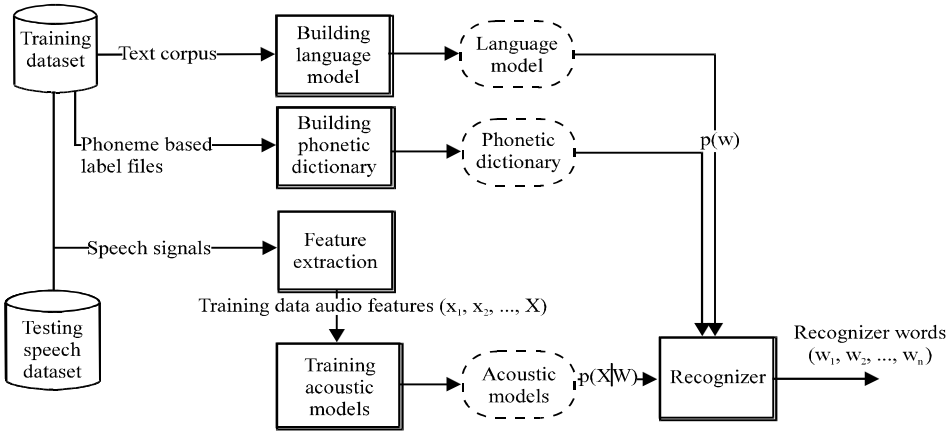

Fig. 1: The architecture of a typical speech recognition system

features that are used to build the HMMs. Also, the features of the testing set are extracted and used as the recognizer input to generate the sequence of uttered words. There are multiple sets of acoustic features that can be used but the most common features set are the Mel Frequency Cepstral Coefficients (MFCCs) (Azim *et al*., 2016a, b). The MFCCs are so, defined that they are "biologically inspired" they resemble the human auditory system which makes them the most appropriate acoustic representation for speech signals in speech recognition systems.

Recognition engine: The recognizer's main goal is to decode the input signal into a sequence of words that's uttered in the input speech signals. It utilizes the built HMMs, phonetic dictionary and language models to accomplish its task. Using such a decoding algorithm (i.e., Viterbi decoding (Forney, 1973; Viterbi, 2006) it can initially map the input file feature vectors into a set of phonemes or tri-phones in case the acoustic models were phoneme based or tri-phones based. After that, it makes use of the phonetic dictionary and the language models to refine the output to a meaningful sentence.

HMM state tying techniques: In order to create a LVCSR we need to guarantee that all the language contexts are trained and can be recognized accurately. To do so, in a traditional way, we have to precisely choose the training data and analyze the gathered data to determine if, there are any missing contexts.

To overcome that problem, state tying techniques where used to reduce the total number of HMM parameters in the training process. State tying (or state clustering) means that the similar states share the same set of parameters. The tying techniques are used to reduce the number of HMM parameters that are need to be estimated. In order to determine which states could be tied together aclustering algorithmis needed.

These can be done using either a bottom up approach or a top down approach. Bottom up approaches assume at the beginning that all the contexts are distinct then the similar models are merged. These approaches use a distance measure to decide which contexts are acoustically similar. The maindrawback of this approach is that they require a training sample for each context to build the initial model.

The other top down approaches use some classification techniques such as k-means and decision trees to classify the contexts into more specific ones. One of the most common top down approach is using a linguistic knowledge to represent all possible phonetic classes that may occur along with the training data to determine which contexts are similar enough to be tied together. Although, this approach is a language dependent approach, it is able to estimate model parameters for all language contexts even, if some have not appeared in the training data.

Data driven based state tying: By measuring the distance between distributions, the data driven state tying approach can decide which contexts are similar and can be tied. Initially all the states are placed in individual clusters the pair of clusters which when combined would form the smallest resultant cluster and are merged. This process sepeats until either the size of the largest cluster reaches the threshold or the total number of clusters have been reached. The distance metric used in this approach is the Euclidean distance in Eq. 3 or the weighted Euclidean distance in Eq .4 and it depends on the type of state distribution. For single Gaussians, a weighted:

$$\left[\sum_{d=1}^{D}\left(\mu_{x_d}-\mu_{y_d}\right)^2\right]^{\frac{1}{2}} \tag{3}$$

$$\left[\sum_{d=1}^{D}\frac{\left(\mu_{x_d}-\mu_{y_d}\right)^2}{\sigma_{x_d}\sigma_{y_d}}\right]^{\frac{1}{2}} \tag{4}$$

where, D is the dimension of feature space, x and y are two clusters with means of μ_{x_d} and μ_{y_d} and standard deviations of σ_{x_d} and σ_{y_d}, respectively. Euclidean distance between the means is used and for tied mixture systems a Euclidean distance between the mixture weights is used. The main drawback of the data driven approach is that it cannot handle the unseen language contexts in other words it cannot recognize the tri-phones that were not trained during the training process and subsequently, no HMMs were generated. On the contrary the tree based state tying technique can handle that problem and is able to find the most suitable output distribution for the unseen tri-phones from the trained models.

Decision tree based state tying: The tree based phonetic decision state tying technique uses the trained models obtained from the training data and phonetic questions to classify the models state. Each of the phonetic questions represent one of the phoneme classes and should include every possible context that can affect the acoustic recognition of the phoneme and should range from the single specific instance of each phoneme to wider general classes (e.g., nasal, fricative, etc). In the tri-phones system the questions should take into consideration both the left

Table 1: Arabic vowels phonemes inventory

Variables	Front	Central	Back
High	i§		uã
Mid			
Low		a~	

Table 2: Arabic consonants phonemes inventory

Variables	Labial	Labio-Dental	Interdental	Dental	Alveolar	Palatal	Velar	Uvular	Pharyngeal	Glotal
Stop										
Voiceless				Tt			k			
Voiced	b			Dd						
Fricative										
Voiceless		F	V	Ss	$		X		H	h
Voiced			Z*	z			g		E	
Affricate										
Voiceless										
Voiced					J					
Glide	w					Y				
Nasal	m				N					
Liquid				I	R					

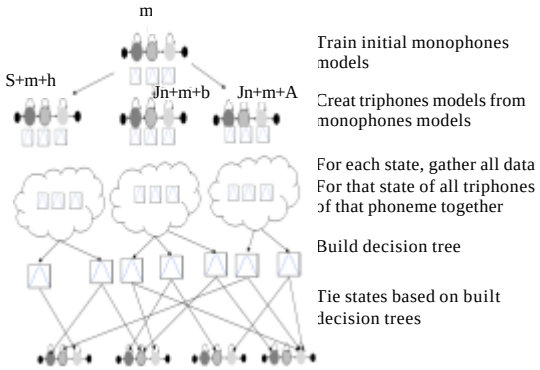

Fig. 2: The detailed process of phonetic decision tree state tying

and the right contexts of the phonemes. The Arabic phonemes inventories for both the vocalic and consonantal phonemes are showed in Table 1 and 2, respectively. Vowels are represented in terms of height and backness of the tongue's position. In Table 1, the rows represent the height of the tongue's position while the columns represent its backness. In Table 2, the rows represent different manners of articulation while the columns represent different placess of articulation (Habash, 2010).

The detailed processes of generating the unseen tri-phones models as described in Fig. 2 goes as follows; first of all, a set of monophonic based models is generated from the training models. After that the tri-phone models are generated from the built monophonic ones by cloning and tying the parameters of the generated monophonic models based on the set of tri-phones given. A state buffer of each state is assumed and all the data from corresponding states of the generated tri-phone models are gathered in that buffer state. Build the decision tree as shown in Fig. 2. Finally, the tri-phones that appear at the same leaf node are considered as one cluster and their values are tied together, if corresponding to a generated model and the unseen tri-phone model parameters are estimated from the parameters pool are considered.

Experiments: Hidden Markov Models Toolkit (HTK) is a portable tool kit for generating and manipulating HMMs. HTK consists of a set of tools used for analyzing speech model training, testing and results analysis. The tools source is available in c language. HTK was originally developed at the machine intelligence laboratory (known as the speech vision and robotics group) of the Cambridge University Engineering Department (CUED) where it has been used to build CUED's large vocabulary speech recognition systems (Young *et al*., 2013; Young and Young, 1993). The Arabic LVCSR acoustic models in this research were trained and tested using HTK tools and the results were analyzed using the Hresults tool in the tool kit (Young *et al*., 2006).

Dataset: A large scale Arabic single speaker corpus (Abdo *et al*., 2014; Jafri *et al*., 2015; Almosallam *et al*., 2013) was used in the implemented experiments. A total of 7 h of recordings represented in 4372 wav files were used. Each wav file is aligned with its text transcription file and the phonemic based alignment file. The selection and the recording of the corpus were done under the supervision of professional linguists. The total number of words in the corpus is 51,432 with 21,566 unique words.

RESULTS AND DISCUSSION

In order to evaluate the performance of the developed system, the Word Correctness ($W_{Corr.}$) and the Word accuracy (W_{ACC}) measures where used in Table 3. The

Table 3: W_{ACC} and $W_{Corr.}$ of the Tied States tri-phones and the monophonic models

	Tied States tri-phones models					
	Tree based		Data driven based		Monophonic models	
No. of Gaussians	$W_{Corr.}$ (%)	W_{ACC} (%)	$W_{Corr.}$ (%)	W_{ACC} (%)	$W_{Corr.}$ (%)	W_{ACC} (%)
1	88.59	61.22	76.97	49.61	49.2	35.4
3	89.04	62.37	76.37	49.67	50.91	36.96
5	87.34	61.36	74.43	48.26	58.41	50.96
9	88.87	63.03	75.40	49.36	58.17	50.66
16	90.82	65.04	76.81	50.64	59.23	51.36
20	93.43	68.28	77.41	51.86	58.45	51.2
25	94.33	69.53	77.97	53.02	58.15	50.65
30	94.99	70.26	77.97	52.84	NA	NA
35	95.06	70.47	78.03	52.96	NA	NA
40	95.13	70.54	77.88	52.87	NA	NA
45	95.06	70.57	77.50	52.90	NA	NA

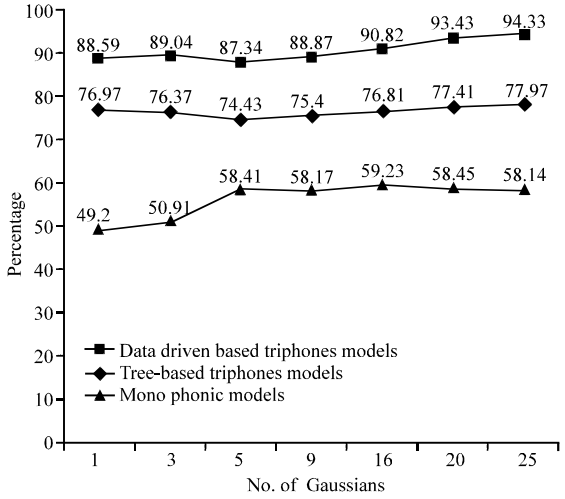

Fig. 3: The word correctness of the monophonic models vs. tri-phonic models

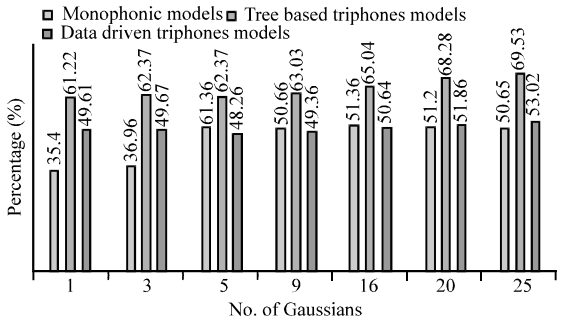

Fig. 4: Monophonic based models vs. tri-phones based models word correctneess (%)

results tool from the HTK toolkit counts the total No. of words (N), word Hits (H), Deletions (D), Insertions (I) and Substitutions (S) using both the original and recognized word based label files. The $W_{Corr.}$ and W_{ACC} are calculated as showed in Eq. 4 and 5, respectively (Fig. 4 and 5):

$$W_{Corr.} = \frac{H}{N} \times 100\% = \frac{N - S - D}{N} \times 100\% \quad (4)$$

$$W_{ACC.} = \frac{H - 1}{N} \times 100\% = \frac{N - S - D - 1}{N} \times 100\% \quad (5)$$

CONCLUSION

This study introduced an ongoing research towards developing a large vocabulary continuous speech recognition system for the Arabic language. The presented approach is based on applying a phonetic decision tree for the tying HMM states in order to recognize unseen tri-phones that may occur that did not appear in the training dataset. First, the monophonic models were built then the tri-phones models were generated. After that, the states of tri-phones models were tied using both approaches of state tying. The experiments were conducted on 7 h of recording and the tree based tied tri-phone models achieved a word correctness of 95.13% using 40 Gaussian mixtures while the data driven based tied models and the phoneme based models achieved a word correctness of 77.97% using 25 Gaussian mixtures and 58.45 using 20 Gaussians mixtures, respectively.

REFERENCES

Abdo, M.S., A.H. Kandil and S.A. Fawzy, 2014. MFC peak based segmentation for continuous Arabic audio signal. Proceedings of the Middle East Conference on Biomedical Engineering (MECBME), February 17, 20, 2014, IEEE, Giza, Egypt, ISBN:978-1-4799-4799-7, pp: 224-227.

Almosallam, I., A. AlKhalifa, A.M. Ghamdi, M.I. Alkanhal and A. Alkhairy, 2013. SASSC: A standard Arabic single speaker corpus. Proceedongs of the ISCA Conference on SSW Synthesis Workshop, August 31-September 1, 2013, ISCA, Barcelona, Spain, pp: 249-253.

Alotaibi, Y.A., 2008. Comparative study of ANN and HMM to Arabic digits recognition systems. J. King Abdulaziz Univ. Eng. Sci., 19: 43-59.

Azim, M.A., A.A.A. Hamid, N.L. Badr and M.F. Tolba, 2016a. Tree-Based HMM state tying for Arabic continuous speech recognition. Proceedings of the International Conference on Advanced Intelligent Systems and Informatics, October 18, 2016, Springer, Berlin, Germany, ISBN:978-3-319-48307-8, pp: 96-103.

Azim, M.A., N.L. Badr and M.F. Tolba, 2016. An enhanced Arabic phonemes classification approach. Proceedings of the 10th International Conference on Informatics and Systems, May 9-11, 2016, ACM, Giza, Egypt, ISBN:978-1-4503-4062-5, pp: 210-214.

Azmi, M.M. and H. Tolba, 2008. Syllable-based automatic Arabic speech recognition in different conditions of noise. Proceedings of the 9th International Conference on Signal Processing ICSP08, October 26-29, 2008, IEEE, Egypt, ISBN:978-1-4244-2178-7, pp: 601-604.

Beulen, K. and H. Ney, 1998. Automatic question generation for decision tree based state tying. Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 2, May 15, 1998, IEEE, Germany, ISBN:0-7803-4428-6, pp: 805-808.

Choubassi, M.M.E., E.H.E. Khoury, C.J. Alagha, J.A. Skaf and A.M.A. Alaoui, 2003. Arabic speech recognition using recurrent neural networks. Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology ISSPIT03, December 17, 2003, IEEE, Beirut, Lebanon, ISBN:0-7803-8292-7, pp: 543-547.

Elhadj, Y.O.M., M. Alghamdi and M. Alkanhal, 2014. Phoneme-Based Recognizer to Assist Reading the Holy Quran. In: Recent Advances in Intelligent Informatics, Thampi, S.M., A. Abraham, S.K. Pal and J.M.C. Rodriguez (Eds.). Springer, Berlin, Germany, ISBN:978-3-319-01777-8, pp: 141-152.

Elshafei, M., 1991. Toward an Arabic text-to-speech system. Arabian J. Sci. Eng., 16: 565-583.

Elshafei, M., H. Almuhtasib and M. Alghamdi, 2002. Techniques for high quality text-to-speech. Inf. Sci., 140: 255-267.

Fahad, A.H. and A. Otaibi, 2001. Speaker-dependant continuous Arabic speech recognition. MSc Thesis, King Saud University, Riyadh, Saudi Arabia.

Farghaly, A. and K. Shaalan, 2009. Arabic natural language processing: challenges and solutions. ACM Trans. Asian Language Inform. Process. Assoc. Comput. Mach., 8: 1-22.

Forney, G.D., 1973. The viterbi algorithm. Proc. IEEE, 61: 268-278.

Habash, N.Y., 2010. Introduction to Arabic natural language processing. Synth. Lectures Hum. Lang. Technol., 3: 1-18.

Hyassat, H. and R.A. Zitar, 2006. Arabic speech recognition using SPHINX engine. Intl. J. Speech Technol., 9: 133-150.

Imperl, B., Z. Kacic, B. Horvat and A. Zgank, 2003. Clustering of triphones using phoneme similarity estimation for the definition of a multilingual set of triphones. Speech Commun., 39: 353-366.

Jafri, A., I. Sobh and A. Alkhairy, 2015. Statistical formant speech synthesis for Arabic. Arabian J. Sci. Eng., 40: 3151-3159.

Lazarides, A., Y. Normandin and R. Kuhn, 1996. Improving decision trees for acoustic modeling. Proceedings of the 4th International Conference on Spoken Language ICSLP 96, Vol. 2, October 3-6, 1996, IEEE, Quebec, Canada, ISBN:0-7803-3555-4, pp: 1053-1056.

Mourtaga, E., A. Sharieh and M. Abdallah, 2007. Speaker independent Quranic recognizer based on maximum likelihood linear regression. Perform. Improv., 316: 61-67.

Nahar, K., A.H. Muhtaseb, A.W. Khatib, M. Elshafei and M. Alghamdi, 2015. Arabic phonemes transcription using data driven approach. Int. Arab J. Inf. Technol., 12: 237-245.

Nahar, K.M., A.W.G. Khatib, M. Elshafei, A.H. Muhtaseb and M.M. Alghamdi, 2013. Data-driven Arabic phoneme recognition using varying number of HMM states. Proceedings of the 2013 1st International Conference on Communications, Signal Processing and their Applications (ICCSPA), February 12-14, 2013, IEEE, Dhahran, Saudi Arabia, ISBN:978-1-4673-2820-3, pp: 1-6.

Nofal, M., A.E. Raheem, E.H. Henawy and N.A. Kader, 2004. Acoustic training system for speaker independent continuous Arabic speech recognition system. Proceedings of the Fourth IEEE International Symposium on Signal Processing and Information Technology, December 18-21, 2004, IEEE, Cairo, Egypt, ISBN:0-7803-8689-2, pp: 200-203.

Odell, J.J., 1995. The use of context in large vocabulary speech recognition. Ph.D Thesis, Cambridge University, Cambridge, England.

Odell, J.J., P.C. Woodland and S.J. Young, 1994. Tree-based state clustering for large vocabulary speech recognition. Proceedings 1994 International Symposium on Speech, Image Processing and Neural Networks ISSIPNN'94, April 13-16, 1994, IEEE, Cambridge, England, ISBN:0-7803-1865-X, pp: 690-693.

Reichl, W. and W. Chou, 1998. Decision tree state tying based on segmental clustering for acoustic modeling. Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 2, May 15, 1998, IEEE, New Jersey, USA., ISBN:0-7803-4428-6, pp: 801-804.

Satori, H., M. Harti and N. Chenfour, 2007. Arabic speech recognition system based on CMUSphinx. Proceedings of the International Symposium on Computational Intelligence and Intelligent Informatics ISCIII'07, March 28-30, 2007, IEEE, Morocco, ISBN:1-4244-1157-2, pp: 31-35.

Viterbi, A.J., 2006. A personal history of the Viterbi algorithm. IEEE. Signal Process. Mag., 23: 120-142.

Young, S., G. Evermann, M. Gales, T. Hain and D. Kershaw *et al*., 2006. The HTK Book (v3.4). Cambridge University, Cambridge, England.

Young, S., P. Woodland, G. Evermann and M. Gales, 2013. The HTK Toolkit 3.4.1. Cambridge University, Cambridge, England,.

Young, S.J. and S. Young, 1993. The HTK hidden Markov model toolkit: Design and philosophy. Ph.D Thesis, Department of Engineering, University of Cambridge, Cambridge, England.

Young, S.J., 1992. The general use of tying in phoneme-based HMM speech recognizers. Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP-92, Vol. 1, March 23-26, 1992, IEEE, Cambridge, Massachusetts, ISBN: 0-7803-0532-9, pp: 569-572.

Zgank, A., B. Horvat and Z. Kacic, 2005. Data-driven generation of phonetic broad classes, based on phoneme confusion matrix similarity. Speech Commun., 47: 379-393.