Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 7.056

IJCSMC, Vol. 10, Issue. 12, December 2021, pg.39 – 45

Automatic Phonemes Segmentation for Quran Verses Using Kaldi Toolkit

Alaa Ehab Sakran[1]; Mohsen Rashwan[2,3]; Sherif Mahdy Abdou[2,4]
[1]PhD. Program in Computer Science, Sudan University of Science and Technology, Khartoum, Sudan
[2]Research & Development International (RDI®), Giza, Egypt
[3]Department of Electronics and Communication Engineering, Cairo University, Giza, Egypt
[4]Department of IT, Faculty of Computers and Information, Cairo University, Giza, Egypt
alaa238.ehab@gmail.com; mrashwan@rdi-eg.com; sabdou@rdi-eg.com

DOI: 10.47760/ijcsmc.2021.v10i12.006

Abstract: In this paper, automatic segmentation system was built using the Kaldi toolkit at phoneme level for Quran verses data set with a total speech corpus of (80 hours) and its corresponding text corpus respectively, with a size of 1100 recorded Quran verses of 100 non-Arab reciters. Initiated with the extraction of Mel Frequency Cepstral Coefficients MFCCs, the proceedings of the building of Language Model LM and Acoustic Model AM training phase continued until the Deep Neural Network DNN level by selecting 770 waves (70 reciters). The testing of the system was done using 220 waves (20 reciters), and concluded with the selection of the development data set which was 280 waves (10 reciters). Comparison was implemented between automatic and manual segmentation, and the results obtained for the test set was 99% and for the Development set was 99% with Time Delay Neural Networks TDNN based acoustic modelling.

Keywords: Automatic phonetic segmentation, ASSS, AM, LM, GMM-HMM, DNN-HMM, KALDI

1- Introduction:

Automatic segmentation for the holly Quran is challenging because of the lexical variety and data sparseness of the Arabic language. Arabic can be considered as one of the most morphologically complex languages. Reducing the entry barrier to build robust Automatic Speech segmentation system (ASSS) for Arabic has been a research concern over the past decade [1][2]. Unlike American English, for example, which has Carnegie Mellon University CMU dictionary, standard KALDI scripts available, the Arabic language has no freely available resource for researchers to start working on ASSS systems. To build a Holly Quran ASSS system, a researcher does not only need to understand the technical details, but also to have the language expertise, which is a barrier for many people. This has been the main motivation for us to release, and share this with the community. Researchers who are interested in building a baseline Arabic ASSS can use it as a reference.

Significant amount of research has been done in Automatic phonetic segmentation which is a technique that defines boundary locations of certain sounds in a given utterance. Its use is required in situations where phone boundaries or limits must be detected for very large corpora. The areas of speech and speaker recognition, speech synthesis and speech coding use segmentation of speech according to its phonetic transcription [3].

It is typically used to form sub-word units for the purpose of concatenative speech synthesis [4][5], and to determine sound boundaries in massive language corpora Also, to train neural network-based speech recognition systems, or in other applications initiated and increasingly motivated by a study of pronunciation variability based on the analysis of the results of phonetic segmentation. A comprehensive analysis of certain sound realizations can also promote the clinical diagnosis of serious diseases that affect speech production, or the analysis of the pronunciation variability in formal or informal language [6].

Multiple methods were used for the purpose of phonetic segmentation of speech; the most significant method being based on is the Hidden Markov Model (HMM) (Donovan, 1996; Ljolje et al., 1996; Nefti and Boffard, 2001; Pellom, 1998; Toledano and Gmez, 2002). Most modern speech recognition systems use HMM which intuitively forms the backbone of the task of speech segmentation, and is well suited for it [3].

The most crucial issue in the field of speech such as ASR, speech synthesis, speech database and speech identification and speaker verification is the segmentation of continuous speech into their corresponding phonemes. The most commonly proposed phoneme-level speech segmentation utilized is either manual segmentation or automatic segmentation techniques. Manual speech segmentation requires an expert / phonetician and its segmentation is based on listening and visual assessment of the boundaries required. However, manual phonetic segmentation is tedious, expensive, inconsistent, error-prone, and time-consuming [17].

Due to what was previously said in respect of manual speech segmentation, the development and need of automatic speech segmentation became increasingly important and needed. In brief, automatic speech segmentation techniques were divided into two types, namely, supervised and unsupervised segmentation techniques. Supervised procedures require prior expert knowledge of phoneme boundaries [15], [16]. These boundaries of the phonemes were in the form of their pre-segmented ones. It also required predefined models of the phoneme set of a particular language. On the other hand, unsupervised methods do not require a predefined model and no previous expert knowledge of phoneme sets or their limits are needed. It is mostly used in automatic speech segmentation through new modelling and training data sets [17].

Hence, for a given utterance with available acoustic implementation and known content, ideally on a phonetic level, the basic solution for determining sound boundaries or limits is based on a forced alignment of trained HMM models. In training acoustic models of speech recognizers, this technique is used by default as an important step. It can be performed using various toolkits that implement HMM-based speech recognition, e.g. HTK [7], Sphinx [8]or KALDI [10]. Above all, KALDI stands as the most popular language research toolkits worldwide.

In 2006, a diversity of a new procedure for learning (DNNs) with many hidden layers of nonlinear units, introduced a paradigm escalation in the area of machine learning and artificial intelligence. DNNs were successfully experimented for acoustic modelling, and have shown outstanding performance. Researches showed confirmed proof that the use of DNN improves the accuracy of speech segmentation. Therefore, the development of DNN-HMM hybrid acoustic models is growing rapidly and on a large scale respectively in all knowledge areas of speech Researches.

2- Acoustic Models of Automatic Speech Segmentation:

The acoustic model models the relationship between the acoustic features and phonemes. Usually, a GMM-HMM or a DNN-HMM model is used here. The lexical model models the relationship between the phonemes and the words. It is a pronouncing dictionary or some grapheme-to-phoneme models. The language model assigns a probability to a sequence of words. High probability indicates that the sequence is more likely to occur in a given language. The language model provides context to distinguish between words and phrases that sound similar. The language model is usually implemented in an N-gram fashion.

2.1- Acoustic Model Based on GMM-HMM:

Acoustic model is the fundamental component of the ASSS. The GMMHMM- based acoustic model achieves its great success in ASSS, since the introduction of the Expectation Maximization (EM) algorithm for joint training of GMMs and HMMs. The GMMs give acoustic feature a score which indicates that the probability of the acoustic feature is generated by a HMM state. Then the scores are used in HMM to decode the input into phonemes. Despite, the success of the GMM-HMM models, GMMs have some shortcomings, especially on the modelling efficiencies, while, a discriminative model, like DNN, can do better [11].

2.2- Acoustic Model Based on DNN-HMM:

Deep Neural Network (DNN) is an artificial neural network with multiple hidden layers between the input and output layers. DNN is a powerful classifier, and showed its strength in the speech recognition [12], object recognition [13], natural language understanding [14], and so forth. In the ASR research, DNN is used as an alternative of the GMM. It assigns scores for each acoustic feature to HMM states. Those scores are then used for HMM decoding. A DNN-based acoustic model is shown in Fig. 1. The input of the DNN is a window of frames of real-valued acoustic coefficients. The output layer is a soft-max layer that contains one unit for each possible state of each HMM. The DNN is trained to predict the HMM state corresponding to the central frame of the input window. These targets are obtained by using a baseline GMM-HMM system to produce a forced alignment [11].

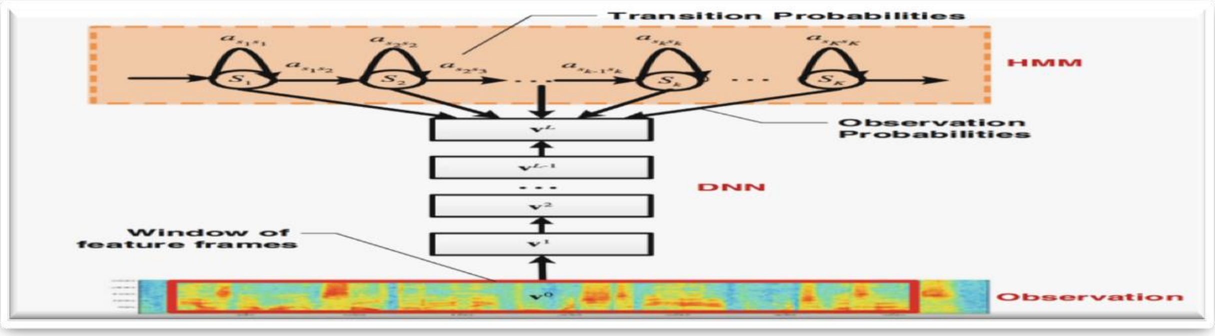

Figure 1:Tthe diagram of a HMM-DNN architecture.

3- Experiments:

The experimental part of this research focused on the analysis of the phonetic segmentation accuracy with data sets recorded by reciters from the holy Quran following surah's respectively (al-Fatiha, al-Asr, al-Kawthar, al-Ikhlas, al-Falaq and an-Nas)

3.1- Experiment methodology:

3.1.1- Automatic Segmentation System Design at Phonetic Level:

The system started with the initialization phase. According to Kaldi toolkit, Speech features were calculated and used in KALDI recipes. As primary Cepstral features (used in AMs mono and tri1), we utilised 13 MFCCs including the zeros Cepstral coefficient, which were calculated for the short-term frame with the length of 25 ms and with the step of 10 ms above the signal have been shifted. The Cepstral mean of normalization was designated to this 13 element vector of short term static features, and delta (dynamic) and delta-delta (acceleration) features accomplished them to the final length of 39. Linear Discriminative Analysis LDA features (used in AM tri2) was evaluated from the context obtained by splicing 5 short term feature vectors on both sides, continuing by LDA and Maximum Likelihood Linear Transform MLLT, which recognises decorrelation and the reduction of the dimension to the length of 40.

After the generation of the feature, all the words in the transcription must be implemented by the creation of a record-specific dictionary. Given that the orthography is to be phonetical, so the words in the lexicon are made up of their graphemes, which non-experts manually fit to the nearest match eventually. Kaldi uses the lexicon, acoustic model, and transcripts to create dataset-specific finite state transducers as a final preparation for the alignment. The baseline 3-gram model was created using SRI Language Modelling Toolkit (SRILM).

In pursuance of AM tri3, a linear regression with maximum probability feature-space Maximum Likelihood Linear Regression (fMLLR) followed in the feature space for every speaker (also called Speaker Adaptive Training SAT). Ultimately, these 40-dimensional fMLLR features with mean and variance normalization in a mutual context were applied as input in this AM, We selected TDNN based acoustic modeling.

The phonetic segmentation depends on the quality of the inputs, obviously, this means the quality of acoustic data, but it also depends heavily on the accuracy of the phonetic content entered. The phonetic content of utterances, which are usually transcribed previously at the orthographic level, can be acquired by a grapheme-to-phoneme conversion or from a pronunciation lexicon, which can also cover pronunciation variability [9] by including more pronunciation variants. This approach must be used when setting sound boundaries for spontaneous and informal speaking, a higher diversity of language dialects and other conditions with relatively high pronunciation variability [16]. It can be held manually (for some very special condition) or automatically (to add certain sound substitutions or reductions to the regular pronunciation based on pre-specified conditions [9], [10]. The DNN training was done using one Graphics Processing Unit GPU on a single machine.

We have observed that using a GPU speeds up training by about a factor of 30 faster than just using the CPU in our setup. Without using a GPU, it would take about three months to train the best system.

3.1.2- Datasets:

Quran verses data set has been recorded with having its corresponding text corpus and number of waves 1100 for 100 reciters and a total speech corpus of (80 hours). All Quran verses text corpus are recorded in mono channel, *.wav format and 16 kHz sample frequency. The dataset was split into a training set, a test set, and development set to simulate the real data collection and training procedure and to avoid having overlap between training, test, and development sets. The data set contains verses from (al-Fatiha, al-Asr, al-Kawthar, al-Ikhlas, al-Falaq and an-Nas) and assuming that the reciters were non-Arab speakers. I also recorded 10 letters which have difficulties in pronunciation to increase the system learning capabilities. The training phase was implemented, by selecting 770 waves for 70 reciters. The testing of the system was done using 220 waves for 20 reciters. Lastly, the development dataset was selected from 10 reciters with a total of 110 waves, all respectively without overlap or repetition.

	Speakers	**Hours**	**Num. boundaries**
Training Set	70	56	-
Test Set	30	16	159,940
Development Set	10	8	79,970

3.1.3- Evaluation Criteria:

When we have two pairs of reference and transcribed boundaries for each phone realization, i.e. $\mathrm{beg}_{ph,ref}[i]$ and $\mathrm{end}_{ph,ref}[i]$ vs.$\mathrm{beg}_{ph}[i]$and $\mathrm{end}_{ph}[i]$, the following two criteria Phone Beginning Error (PBE) and Phone End Error (PEE) can be defined as

$$PBE_{ph}[i] = |beg_{ph}[i] - beg_{ph,ref}[i]|, \quad (1)$$

$$PEE_{ph}[i]=|end_{ph}[i] - end_{ph,ref}[i]|. \quad (2)$$

The accuracy of phone boundary can be approximated using the rate of phone boundary error which is below the chosen threshold which can be defined as:

$$PBE_{ph,thr} = \frac{\sum_{i=0}^{Nph}(PBEph[i] < thr)}{Nph} \qquad (3)$$

Where ph is phone identification, N_{ph} is the number of phone realizations, and thr is the value of the chosen error threshold. Similarly, the same procedure is applied for the computation of $PEE_{ph,thr}$ [18] Threshold values used for realized evaluations within this work were 5 or 10ms respectively.

Results:

In order to measure the performance of automatic speech segmentation, phoneme mapping concept is used. The manually segmented phoneme boundaries are compared with each phoneme sequences found during automatic speech segmentation. The formulas used to calculate the accuracy of the Automatic phonetic segmentations mention in the Evaluation Criteria section, in our study the accuracy was evaluated using two different tolerance values 5ms, 10ms as can be seen in table 1.

	% performance for ≥ 5ms	**% performance for ≥ 10ms**
Test Set	98 %	99 %
Dev Set	98 %	99 %

Table 1: The obtained results for automatic phonetic segmentation accuracy, 98% for a 5ms and 99% for a 10ms

Conclusion and Future Works:

In this paper we have studied DNN-based AM for holy Quran verses, using the Kaldi toolkit. We have experimented with DNNs with different numbers of hidden layers, the automatic phonetic segmentation experiments showed that the results obtained was 99 % compared with manual segmentation (tolerance value was 10ms). In further research, we will investigate in the holy Quran by using some other DNN's configurations and types.

Acknowledgements:

We would like to thank RDI Company based in Egypt for the facilitations it provided and for its continuous support.

References

[1]. F. Diehl, M. J. F. Gales, M. Tomalin, and P. C. Woodland,m " Morphological decomposition in Arabic ASR systems," *Comput. Speech Lang.*, vol. 26, no. 4, pp. 229–243, Aug. 2012.

[2]. B. Kingsbury, H. Soltau, G. Saon, S. Chu, H.-K. Kuo, L. Mangu, S. Ravuri, N. Morgan, and A. Janin, "The IBM 2009 GALE Arabic speech transcription system," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 4672–4675

[3]. Dines, John, Sridha Sridharan, and Miles Moody. "Automatic speech segmentation with hmm." In Proceedings of the 9th Australian Conference on Speech Science and Technology, pp. 544-549. 2002.

[4]. Matouˇsek, J., Tihelka, D., Psutka, J.: Experiments with automatic segmentation for Czech speech synthesis. In: Matouˇsek, V., Mautner, P. (eds.) TSD 2003. LNCS (LNAI), vol. 2807, pp. 287–294. Springer, Heidelberg (2003). https://doi.org/10.1007/978-3-540-39398-6 41

[5]. Rendel, A., Sorin, A., Hoory, R., Breen, A.: Toward automatic phonetic segmentation for TTS. In: Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, pp. 4533–4536 (2012)

[6]. Mizera, P., Pollak, P., Kolman, A., Ernestus, M.: Impact of irregular pronunciation on phonetic segmentation of Nijmegen corpus of casual Czech. In: Sojka, P., Hor´ak, A., Kopeˇcek, I., Pala, K. (eds.) TSD 2014. LNCS (LNAI), vol. 8655, pp. 499–506. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10816-2 60

[7]. Young, S., et al.: The HTK Book, Version 3.4.1. Cambridge (2009)

[8]. CMUSphinx: Open source speech recognition toolkit. http://cmusphinx.github.io

[9]. Brunet, R.G., Murthy, H.A.: Pronunciation variation across different dialects for English: a syllable-centric approach. In: 2012 National Conference on Communications (NCC) (2012)

[10]. Povey, D., et al.: The Kaldi speech recognition toolkit. In: Proceedings of the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, ASRU 2011 (2011)

[11]. Zhang, Hui, Feilong Bao, and Guanglai Gao. "Mongolian speech recognition based on deep neural networks." In Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data, pp. 180-188. Springer, Cham, 2015.

[12]. G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[13]. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[14]. Y. Bengio, H. Schwenk, J.-S. Sen´ecal, F. Morin, and J.-L. Gauvain, "Neural probabilistic language models," *Innovations in Machine Learning*, pp. 137–186, 2006.

[15]. F. D. Brugnara F., and Omologo M., "Automatic segmentation and labeling of speech based on hidden Markov models," *Speech Commun. ,* vol. 12, pp. 357-370, 1993.

[16]. B. L. Pellom, and Hansen, J. H. L., "Automatic segmentation of speech recorded in unknown noisy channel characteristics," *Speech Commun.25,* pp. 97-116, 1998.

[17]. Y. C. a. Q. Wang, "A Speaker Based Unsupervised Speech Segmentation Algorithm Used in Conversational Speech," *Springer-Verlag Berlin Heidelberg 2007,* pp. 396–402, 2007.

[18]. Khan, Arif, and Ingmar Steiner. "Qualitative Evaluation and Error Analysis of Phonetic Segmentation." Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2017 (2017): 138-144.