AUTOMATIC ARABIC PRONUNCIATION SCORING FOR LANGUAGE INSTRUCTION

Hassan Dahan, Abdul Hussin, Zaidi Razak, Mourad Odelha

University of Malaya (MALAYSIA)
hasbri@um.edu.my

Abstract

Automatic articulation scoring makes the computer able to give feedback on the quality of pronunciation and eventually detect some phonemes miss-pronunciation. Computer-assisted language learning has evolved from simple interactive software that access the learner's knowledge in grammar and vocabulary to more advanced systems that accept speech input as a result of the recent development of speech recognition[1]. Therefore many computer based self teaching systems have been developed for several languages such English, Deutsch and Chinese, however for Arabic; the research is still in its beginning. This study is part of the "Arabic Pronunciation improvement system for Malaysian Teachers of Arabic language" project which aimed at developing computer based systems for standards Arabic language instruction for Malaysian teachers of Arabic language. The system aims to help teachers to learn Arabic language quickly by focusing on the listening and speaking comprehension (receptive skills) to improve their pronunciation[2,3]. In this paper we addressed the problem of giving marks for Arabic pronunciation by using a Automatic Speech Recognizer (ASR) based on Hidden Markov Models (HMM), thus our approach to pronunciation scoring is based on the HMM log-likelihood probability.

Keywords: Arabic Pronunciation scoring, Hidden Markov Models (HMMs), Log-Likelihood probability, Baum Welch algorithm, Viterbi algorithm.

1 INTRODUCTION

In order to ameliorate interaction between humans and machines, much research in automatic speech recognition by computer has been done during last decades. In fact, speech has the potential to be a better interface than other computing devices used such as keyboard or mouse [4, 5]. The word online here means that the user will use the microphone to speak one word before the system will score the pronunciation by giving a mark. As speech database files are relatively available for English language comparing to other languages such Malay [6, 7, 8] thus the English language will be first used to develop and test the system. Then other languages will be used especially Malay and Arabic, as it is sufficient to record a speech database of Malay or get it from internet (if available) to extend the system. This project is based on Hidden Markov Model (HMM) which widely used technique in pattern classification and especially in speech recognition. Here, we take benefit of HMM which is very consistent approach for speech recognition.

Hidden Markov Model (HMM) is a natural and highly robust statistical methodology for automatic speech recognition [9]. Thus state-of-the-art speech recognition systems are usually based on the use of HMMs. It is also being tested and proved considerably in a wide range of applications. The model parameters of the HMM are essence in describing the behavior of the utterance of the speech segments. Many successful heuristic algorithms are developed to optimize the model parameters in order to best describe the trained observation sequences; however the Baum Welch (BW) algorithm has demonstrated very high performance that's why it is actually almost the only method used for HMM parameters estimation [10, 11].

2 PROBLEM STATEMENT

There are a lot of people want to learn Arabic language as part of their daily usage, for their future careers and to expand their knowledge. But there are some difficulties that face these learners. These difficulties are how to pronounce Arabic letters and sounds properly and the pronunciation is more confusing because there are strange Arabic sounds that may not be found in the native language of the learner.

Automatic scoring system plays a very important role in the life. It can be used to improve learning of foreign speakers and improve the correct rate. Also, the system will be used to evaluate pronunciation quality of speakers and to communicate between people.

Arabic speech processing research is actually on its beginning, thus there is no research or system found for pronunciation scoring of Arabic language. Therefore research is needed.

3 OBJECTIVES

- To design an Arabic speech pronunciation scoring system.
- To develop and implement the Arabic speech pronunciation scoring system.
- To evaluate and analyze the performance of the Arabic speech pronunciation scoring system.

4 PATTERN RECOGNIZER (HMM TOOL)

In this paper, the Hidden Markov Model (HMM) is used as a statistical method of characterizing and matching the spectral patterns of speech. Baum-Welch algorithm is used to determine the reference pattern that best matches the input feature vectors by comparing stochastic possibility scores while Viterbi algorithm is used to test the HMM set in order to find the optimal state path for an observation sequence, and finally calculate the log-likelihood in which is based the pronunciation scoring.

4.1 HMM training systems

HMM Training is the process of HMM's parameters calculation [12]. As we can see in figure 1, the training tools use the speech data and their transcriptions to estimate the HMM parameters then the recognizer will use these HMMs to classify the unknown speech utterances.

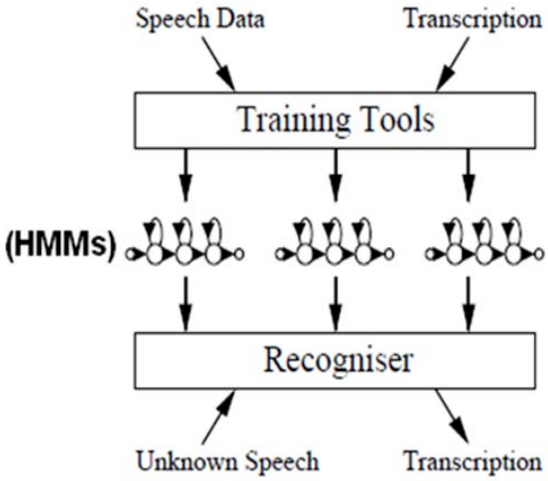

Figure 1. HMM training process [12]

The model parameters of an HMM can be expressed as a set of three elements: $\lambda = \{A, B, \pi\}$ [11]. Where:

- A= { aij } is the state transition probability matrix, each element aij represent the probability that the HMM model will transit from stat I to stat J. Elements of matrix A must satisfy the next two conditions:

$a_{ij} \geq 0$ where $1 \leq i,j \leq 3$ (1)

$$\sum_{j=0}^{3} a_{ij} = 1 \quad \text{where } 1 \leq i \leq 3 \qquad (2)$$

- B ={bij (k)} the observation probability matrix, such that bij is the probability that the observation Ok has been generated by stat i. Elements of matrix B must satisfy the next two conditions:

$b_{ij} \geq 0$ where $1 \leq i,j \leq 3$ (3)

$$\sum_{j=0}^{3} b_{ij} = 1 \quad \text{where } 1 \leq i \leq 3 \quad (4)$$

- π = {πi} the initial state distribution vector, and every πi express the probability that the HMM chain will start at state i. Elements of vector π must satisfy the next two conditions:

$\pi_i \geq 0$ where $1 \leq I \leq 3$ (5)

$$\sum_{i=0}^{3} \pi_i = 1 \quad (6)$$

4.2 System organization

In this work we are realizing a phonemes pronunciation scorer system for Arabic, in order to develop an Arabic pronunciation evaluation and correction tool for non Arabic speakers and especially for Malaysian Arabic language teachers. To detect pronunciation we need to get score for each word in the speech utterance, this leads us to build a set of continuous tied-state triphones HMMs and use them to calculate the log-likelihood of the word and finally calculate the articulation score.

As we can see in figure 2 and 3 Features vectors are extracted from speech utterances, using Mel Frequency Cepstral Coefficients (MFCC) technique, then Baum Welch Algorithm is applied in order to train the system and built the HMMs set. And then in the pronunciation scoring system the HMMs are used with the test speech features to perform a forced alignment of the speech utterances by applying the Viterbi Algorithm, and finally Phonemes pronunciation scores are calculated.

4.2.1 Analysis Speech

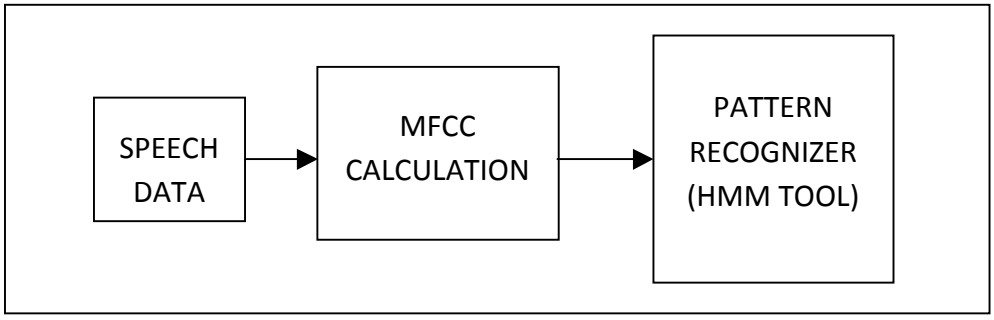

Figure 2. Analysis Speech Block Diagram [12]

4.2.2 MFCC Calculation

Me-Frequency Cepstral Coefficients (MFCC) is a popular set of features often utilized for ASR. MFCC calculation is used in the feature extraction process.

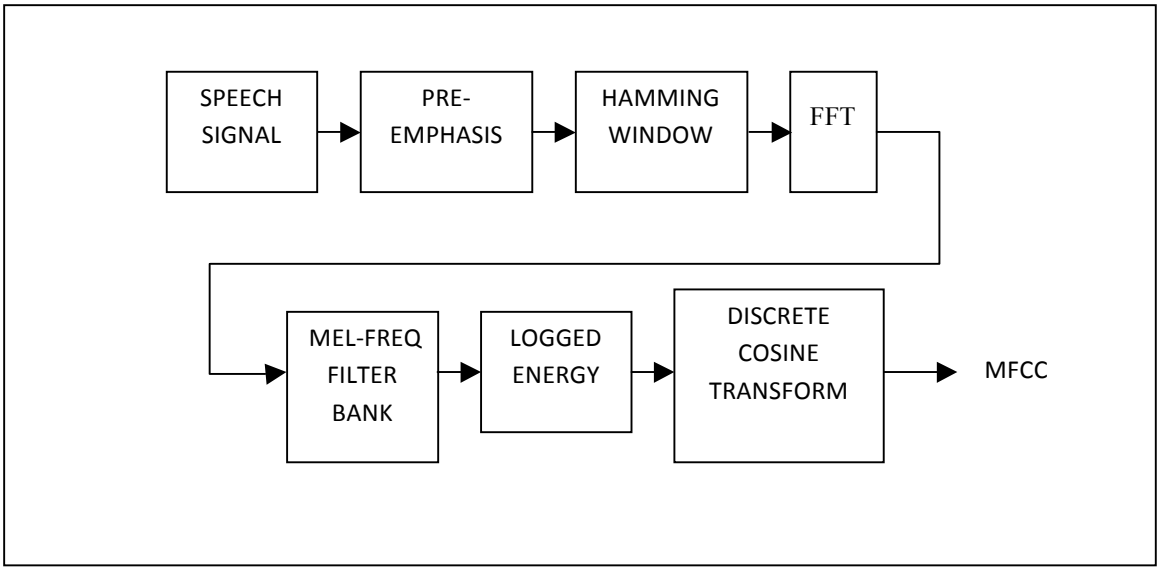

Figure 3. MFCC Calculation Block Diagram [12]

MFCCs are derived step by step as follows:

1. The speech signal is separated into short segments called frame through pre-emphasis process.
2. Multiply each frame using a hamming window in order to keep the continuity of the first and the last point in the frame.
3. Take the Fourier transform of (a windowed extract of) a signal to obtain the frequency magnitude respectively to each frame.
4. A Mel filter bank analysis is performed to obtain a perceptually significant spectral estimate. Map the powers of the spectrum obtained onto the Mel scale, using triangular overlapping windows.
5. For better performance, the logs of the powers at each of the Mel frequencies have been obtained.
6. Discrete Cosine Transform convert the frequency domain into a time-like domain called frequency domain. Take the discrete cosine transform of the list of Mel log powers, as if it were a signal.
7. The amplitudes of the resulting spectrum are similar to cepstrum, therefore it is referred to as the mel-scale cepstral coefficients, or MFCC.

This MFCC will be passed to Hidden Markov Model (HMM) Tool which act as the pattern recognizer by estimating the probability of each phoneme at contiguous, small regions (frames) of the speech signal [13].

4.2.3 HMM training using BW algorithm

The BW algorithm has been used to train the HMM. An initial guess of the HMM parameters is made, after that the BW algorithm is run for 20 iterations to get more accurate parameters. As result we get a continuous density mixture Gaussian HMM. Finally transcriptions of unknown speech utterances will be made by the recognizer module to determine how accurate are the HMM's parameters

4.2.4 Viterbi Algorithm

The Viterbi algorithm is a dynamic programming algorithm for finding the most likely sequence of hidden states called the Viterbi path that results in a sequence of observed events[14, 15], especially in the context of Markov information sources, and more generally, hidden Markov models. The forward algorithm is a closely related algorithm for computing the probability of a sequence of observed events. These algorithms belong to the realm of information theory [16].

The log likelihood [12] has been used, and it represents the probability that the training observation utterances have been generated by the current model parameters and it is a function of the following form [2]:

$$P_n = (\sum_{k=1}^{M} \log (P(O_k/\lambda_n)))/M \quad (7)$$

5 EXPERIMENTS

We have extensively used a pronunciation scoring paradigm for the automatic assessment of pronunciation quality by machine. In this scoring paradigm, both native and nonnative speech data are collected, and a database of human-expert ratings is created. The speech database design is very important, especially for text-independent pronunciation evaluation. Similarly, the reliability of the human ratings is critical, since we see pronunciation evaluation as a prediction problem, where we are trying to predict the grade a human expert would assign to a particular skill by using statistical models constructed with the speech and the expert-ratings databases.

6 RESULTS

In this work, four human scorers have been used to assess the Arabic pronunciation, and their results have been compared to the proposed system, as it is shown in Table 2. In table 1, the percentage of scores given by the human scorers is presented.

Score	1	2	3	4	5
Percentage (%)	7	21	47	20	5

Table 1: Histogram of scores

Human scorer	1	2	3	4	average
System accuracy	90.2	88.32	87.63	92.37	89,63%

Table 2: system results

7 CONCLUSION

This paper proposes a method for pronunciation scoring, which is independent from the student's first language and can in principle be applied to other target languages. Besides investigating features and methods for scoring words and sentences, an approach to automatic diagnosis of phoneme mispronunciations based on word scoring results is presented.

REFERENCES

[1] Pellom, B. & Hacioglu, K. 2001, 'Sonic: The university of Colorado continuous speech recognition system', University of Colorado, Technical Report TR-CSLR-2001-01.

[2] Tabbal, H., El Falou, W. & Monla, B. 2006, 'Analysis and implementation of a "Quranic" verses delimitation system in audio files using speech recognition techniques', Information and Communication Technologies, 2006. ICTTA 'O6. 2nd on 24-28 April 2006, vol. 2, pp. 2979-2984.

[3] Young, S., Ollason, D., Valtchev, V. & Woodland, P. 2002, The HTK Book (for HTK Version 3.2), Cambridge University Engineering Department, Cambridge, England.

[4] Tan, Z. H., Lindberg, B. & Singh, S. 2008, Automatic Speech Recognition on Mobile Devices and over Communication Network, Springer, London.

[5] Deligne, S., Dharanipragada, S., Gopinath, R., Maison, B., Olsen, P. & Printz, H. 2002, 'A robust high-accuracy speech recognition system for mobile applications', In IEEE Transactions on Speech and Audio Processing, vol. 10, no. 8, pp. 551-561.

[6] Motiwalla, L. F. & Qin, J. 2007, 'Enhancing Mobile Learning Using Speech Recognition Technologies: A Case Study', Eighth World Congress on the Management of eBusiness (WCMeB 2007), pp. 18.

[7] Tan, C. L. & Jantan, A. 2004, 'DIGIT RECOGNITION USING NEURAL NETWORKS', Malaysian Journal of Computer Science, vol. 17 ,no. 2, pp. 40-54.

[8] Anil, K. J., Robert, P. W. & Jianchang, M. 2000, 'Statistical Pattern Recognition: A Review', IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 1.

[9] Rabiner, L. & Juang, B. H. 1993. Fundamentals of Speech Recognition. Prentice Hall, Englewood Cliffs, New Jersey.

[10] Tóth, L., Kocsor, A. & Csirik, J. 2005, 'On Naive Bayes in Speech Recognition', International Journal of Applied Mathematics and Computer Science, vol. 15, no. 2, pp. 287–294.

[11] Huang, X., Alex, A. & Hon, H. W. 2001, Spoken Language Processing: A Guide to Theory, Algorithm and System Development. Prentice Hall, Upper Saddle River, New Jersey.

[12] Zaykovskiy, D. 2006, 'Survey of the speech recognition techniques for mobile devices'. In International Conference Speech and Computer, St.Petersburg, Russia, June 2006, pp. 88–93.

[13] Zaykovskiy, D. & Schmitt, A. 2007, 'Java (J2ME) Front-End for Distributed Speech Recognition', 21st International Conference on Advanced Information Networking and Applications Workshops (AINAW'07).

[14] Cohen, J. 2008, 'Embedded Speech Recognition Applications in Mobile Phones: Status, Trends, and Challenges', Acoustics, Speech and Signal Processing 2008, ICASSP 2008, IEEE International Conference, pp. 5352-5355.

[15] Delphin-Poulat, L. 2004, 'Robust speech recognition techniques evaluation for telephony server based in-car applications' Acoustics, Speech, and Signal Processing 2004, Proceedings, (ICASSP '04), IEEE International Conference, vol. 1, pp. I-65-8.

[16] Rabiner, L. R. 1989, 'A tutorial on hidden Markov models and selected applications in speech recognition', Proceedings of the IEEE, vol. 77, no. 2, pp. 257-286.