I.J. Image, Graphics and Signal Processing, **2018, 4, 31-38**
Published Online April 2018 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2018.04.04

A Dataset for Speech Recognition to Support Arabic Phoneme Pronunciation

Moner N. M. Arafa[1,*], Reda Elbarougy[2], A. A. Ewees[1], G. M. Behery[2]
[1] Department of Preparing Computer Teacher, Faculty of Specific Education, Damietta University, Egypt
[2] Department of Mathematics, Faculty of Science, Damietta University, Egypt
* Email: moner.nasef5@gmail.com

Received: 17 October 2017; Accepted: 15 February 2018; Published: 08 April 2018

Abstract—It is difficult for some children to pronounce some phonemes such as vowels. In order to improve their pronunciation, this can be done by a human being such as teacher or parents. However, it is difficult to discover the error in the pronunciation without talking with each student individually. With a large number of students in classes nowadays, it is difficult for teachers to communicate with students separately. Therefore, this study proposes an automatic speech recognition system which has the capacity to detect the incorrect phoneme pronunciation. This system can automatically support children to improve their pronunciation by directly asking children to pronounce a phoneme and the system can tell them if it is correct or not. In the future, the system can give them the correct pronunciation and let them practise until they get the correct pronunciation. In order to construct this system, an experiment was done to collect the speech database. In this experiment 89, elementary school children were asked to produce 28 Arabic phonemes 10 times. The collected database contains 890 utterances for each phoneme. For each utterance, fundamental frequency f0, the first 4 formants are extracted and 13 MFCC co-efficients were extracted for each frame of the speech signal. Then 7 statics were applied for each signal. These statics are (max, min, range, mean, mead, variance and standard divination) therefore for each utterance to have 91 features. The second step is to evaluate if the phoneme is correctly pronounced or not using human subjects. In addition, there are six classifiers applied to detect if the phoneme is correctly pronounced or not by using the extracted acoustic features. The experimental results reveal that the proposed method is effective for detecting the miss pronounced phoneme ("ĺ").

Index Terms—Phoneme pronunciation, Arabic phoneme, Dataset, Arabic Dataset, Reading difficulties, Arabic phoneme pronunciation, feature extraction, MFCC.

I. Introduction

Most natural way of communication between human beings and understanding is the speech [1], we learn speaking before learning read or write, where computers play an essential role in people's lives, it's no surprise exert maximum work to improve these computers so that they are able to distinguish human speech and human response to his orders. Many of computer science researchers do their best to automate this issue by introducing a new trend; it is called speech recognition or sometimes referred to as automatic speech recognition (ASR). Speech recognition is the method of automatically identifying a specific word spoken by a speaker based on individual features included in speech waves to allow the speaker to supply input to a system with your voice or provide input by talking[2][3]. It also allows a machine to distinguish spoken words [2].

Since pronunciation is a foundation in the process of social communication, speech defects are widespread among primary school children verbal defects have a role in the formation of self-concept committed by the child, and thus in his social inclination. Hence the importance of focusing on these problems and the importance of detecting mistakes in the pronunciation of letters in order to help our children develop psychologically and properly adjust.

However, ASR systems were used to let computer recognize what spaeker said and gives the overall recognition rate. However; if the speaker of these systems has a difficulty to pronounce some phoneme correctly the recognition rate will be very low. In order to improve the recognition rate, it is necessary to check the ability of the user of ASR system to correctly pronounce all phonemes. Therefore; the purpose of the study is to construct a system which has the ability to detect the miss pronounced phoneme this system will be called Automatic Pronunciation Error Recognition (APER).

APER and ASR are alternatives to traditional techniques of reacting with a computer, like literal entry by a keyboard. An effective system can replace, or reduce the reliability on, standard keyboard and mouse input. This can especially assist the following: people who have little keyboard skills or experiences, or who are slow typists; as well as, people with physical disabilities that affect either their inserted data or capacity to read what they have inscribed, dyslexic people, or others who have problems with a word or phoneme use and manipulate in a textual format [3].

The problem of ASR belongs to a much larger topic in engineering and scientific tagged pattern recognition. The goal of pattern recognition is to classify objects of interest

into one of the several categories. The targets of interest are generically called patterns and in our case, which are sequences of acoustic vectors that are extracted from an input speech by techniques. Since the classification method in this state is used on features extracted [4].

Error detection methods utilize extracted features where transform speech signal to a best fitting form, reduces the data rate and extract related information from a speech signal. It produces a huge number of features so we need to decrease dimensions for input speech [1] [5]; that can be done by extracting distinct features for speech, besides minimum loss of valuable information, These features bring phonemes characteristics which differ from one speaker to another, so these features perform major role in any system [6][7]. There are three primary types of front-end processing methods or techniques, namely Linear Predictive Coding (LPC), Perceptual Linear Prediction (PLP), and Mel-Frequency Cepstral Coefficients (MFCC), where the latter two are most usually used in state-of-the-art ASR systems [8].

Automatic pronunciation error detection techniques and ASR are used widely in many fields in our life, such as communication, security, health-care, and education. The educational system may be suffering from various problems in speech issues; for instance, some primary students have problems for uttering some phonemes, words, or letter exit (MAKHARIG) or (Phonemes) that due to some reading disorder problems [9]. These problems can affect the education system, especially in Arabic language where could rise when forwarding on standard Arabic letters position, these are shown in Figure 1. The Arabic language has more attention in recent years because it is a unique language with complex morphology nature and structures [10] [11] [12]; so Arabic ASR is a challenging task [13].

Fig.1. The articulation of Arabic letters pronunciation.

Therefore, this paper proposes an ASR mechanism to improve the Arabic speech recognition accuracy by detecting the words that are uttered incorrectly from students.

The main aim is identifying perfect pronunciation for Arabic phonemes. The system can recognize the correct utterance, which has been trained with a number of students, by comparing a speech signal from known students to a database of an unknown student.

By looking at the previous studies that used speech recognition techniques to automatic dealing with speech disorders, skills improvement, dealing with difficulties, and enriching e-learning systems, it turns out that these studies produced fruitful results and promote the educational process especially identifying correct Arabic phonemes utterance; therefore, let's look for taking the advantages of these studies and trying to start from where others have ended. In the following, some of these studies are listed. In [14] a system was developed to aid pronunciation for teaching Arabic and Qur'an recitation rules learning, it was called HAFFSS 1.0. It proved a reliability and accurate performance. The authors of [15] built a voice database using Egyptian Arabic slang by recording 20 hours of Arabic audio dialogues with an Egyptian accent. A Dictionary was organized mainly based on the basic lexical pronunciation of each word. It contained selected audio units commensurate with the Arabic language and audios private Egyptian dialect class. It was also composed a linguistic statistical model using text strings associated with the base audio data. And they introduced a statistical model based on Hidden Markov Model (HMM). Speech recognition experiments result displayed good correct recognition accuracy.

In addition, [16] used ASR technique to develop rehabilitation systems with Egyptian Arabic accent, this system was applied to two tracks. The first track was concerned with an adult suffering from receptive dysphasia disorder and the second was concerned with children having language speech delayed disorder, and monitor their progress.

The authors in [17] used three feature extraction mechanism namely LPC, MFCC, and Real Cepstrum, the results showed that the MFCC achieved more discriminative than the others to examine classifiers by Neural Network (NN) and HMM, the best ensemble-based architecture gives 94.3% correct classification while the best modular-based results in 79.3% for the testing data. Also, [18] [19] evaluated recognition performance of the speech features at different stages of the system. The MFCC was used as a features extraction mechanism. [18] Scheme was evaluated using a speaker independent Arabic database used in HTK, error detection achieves the overall correct rate of speech recognition ranges from 72.73% to 98.19% from different stages and conditions of the system. MFCC also was used by [20] as a feature extraction technique reduced by using Principal Component Analysis (PCA) and Artificial Neural Networks (ANN) for speech recognition. The cognitive system contains a cognitive content represented in an attractive multi-media program which is supported by pictures, songs, practices, and some stories.

The authors in [21] Extracted 13 values for each of these from a data-set 46 speech samples containing the Hindi vowels and trained them using a combined model of SVM and neural network classification to determine their gender using stacking. The results obtained showed the accuracy of 93.48% after taking into consideration the first Mel coefficient.

In [22] (MFCCs) is used to accurately represent the shape of the vocal tract that manifests itself in the envelope of the short time power spectrum. Author in [23] combined features using Formant Frequencies (FF), combined frequency warping and feature normalization techniques using Linear Predictive Coding (LPC) and Cepstral Mean Normalization (CMN). In [24] presents an approach of ASR system based on isolated word structure, features of speech are extracted using MFCC's, (DTW) applied for speech feature matching and KNN techniques employed as a classifier.

The rest of this paper is organized as follow: Section 2 discusses the material and methods. The proposed system is introduced in Section 3. Experiment Results are displayed in Section 4. Section 5 presents discussions. The conclusion is given in the last section.

II. Material and Methods

Mel-Frequency Cepstral Co-efficients (MFCC): Is one of the signal processing techniques that is used in feature extraction when applied to speech sound, extract the 13 co-efficients. This set of 13 MFCC co-efficients are used as a feature vector in the classification task. Where MFCC is derived from a type of cepstral representation of the audio clip, the frequency bands are positioned logarithmically (on the MEL scale), which approaches the human auditory system's response more nearly than the linearly-spaced frequency bands acquired instantly from the DCT or FFT.

This can allow for better processing of data, where it considers the most widely common algorithm that used for recognition system. MFCC's depends on the identified difference of human ear's critical bandwidths with frequency; filters spaced linearly at low frequencies and logarithmically at high frequencies which have been used to capture the phonetically major characteristics of speech, getting the acoustic characteristics of the speech signal is indicated to as extracted feature that is used in both training and recognition phases. It comprises the following steps: Frame blocking, windowing, FFT (Fast Fourier Transform), Mel-Frequency Wrapping and Cepstrum (MFCC) [8], passing some steps like Figure 2 as in below.

The input speech signal is segmented into frames of 20~30 m_s, an overlap is 160 points, then the frame rate is 16000/(320-160) = 100 frames per second [25], where speech signal s(n) is sent to a high-pass filter as like in (1):

$$s_2(n)=s(n)-a*s(n-1) \quad (1)$$

Where s_2 (n) is the output signal and the value of (a) is usually between 0.9 and 1.0. The z-transform of the filter is (2):

$$H(z)=1-a*z^{-1} \quad (2)$$

To compensate the high-frequency part that was suppressed through speaker sound production. Besides, it can also amplify the value of high-frequency formants. The next step is (3):

$$W_n(m)=0.54-0.45\cos(2\pi n/(N_m-1)) \quad 0\le m\le N_m-1 \quad (3)$$

Where Nm stands for quantity of samples within every frame, the output after windowing the signal will be presented as Y (m) = X (m) W_n (m), 0 ≤ m ≤ Nm -1 where Y (m) represents the output signal after multiplying the input signal represented as (X_m) and Hamming window represented by W_n(m). The step number (4) is:

$$D_k = \sum_{m=0}^{N_m-1} D_m e^{\frac{-j\,2\pi km}{N_m}} \quad (4)$$

Where k= 0, 1, 2 ….. Nm-1

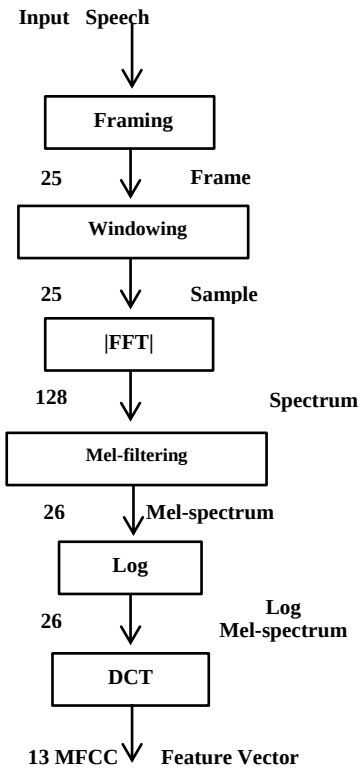

Fig.2. The steps for extracting Mel-frequency cepstral coefficients (MFCC)

FFT is used for doing the conversion from the spatial domain to the frequency domain. Each frame having Nm samples are converted into a frequency domain. Fourier transformation is a fast algorithm to apply Discrete Fourier Transform (DFT), on the given set of N_m samples, passing to (5):

$$f_{mel}=2595\times\log_{10}\left(1+\frac{f}{700}\right) \quad (5)$$

As the frequency gets higher. These filters also get wider. For Mel- scaling mapping is a need to do among the given real frequency scales (Hz) and the perceived frequency scale (Mel's). During the mapping, when a given frequency value is up to 1000Hz the Mel-frequency scaling is linear frequency spacing [15], meeting at last to (6):

$$C_n = \Sigma_{K-1}^{K} (\log D_k) \cos[m(k-\frac{1}{2})\frac{\pi}{k}] \quad (6)$$

Where m = 0, 1… k-1

For converting the log Mel spectrum back into the spatial domain. This transformation either DCT or DFT, both can be used for calculating Co-efficients from the given log Mel spectrum [25].

III. The Proposed Method

The proposed system is designed to solve many of the problems of Arabic phonemes pronunciation where move some steps like:

3.1 Database Collection:

This section represents the method of collecting data and features, where the importance of automatic pronunciation error detection techniques is produced to feed a perfect database. The dataset is represented on audio files where the speech of 89 students (43 male and 46 female) were annotated on segmental pronunciation errors by expert listeners, in the interest of error detector performance.

The descriptions of these steps are explained in Figure 2 and the following subsections.

Preparing the dataset:

A Student Recording: In order to facilitate the testing of the recognizer, improving Arabic phoneme pronunciation mistake identifier, speech database is required. A variety of speech samples were obtained from different speakers (students) to form the speech database. The collected database includes 890 speech samples from 89 different speakers 43 males and 46 females. The digits may be spoken clearly so that it avoids general variations and confusions. The speech is recorded using COOLEDIT software in WAV files have a 16KHz sample rate, MONO channel and 16-bit resolution. The vocabulary items consists isolated Arabic words (letters) of 28 words repeated 10 recorded as all speakers (students). Starting assembly database (audio files), by recording different letters sounds by uttering it from students by pointing to the letter waiting to pronounce it, also by uttering all characters separate to each student, and then for small audio files (Each character in a file).

B Pre-processing: By reviewing audio files separated one by one to include only character pronunciation without additions, taking into consideration the position of a silence at the beginning and end of the section (And the importance of this in the next steps for improving recognition). It is obvious when the voice recording is executed at the school or in the class, there must be noise around or in the recording environment and recognition difficulties, one of the factors have negative affecting speech recognition is negative noise. And percentage and influence in the audio file, where an initial noise-removal work is preferred by adding one or extra filters to noise purification without affecting the audio file or minimizing as it is possible, effect on each audio file, It can be called noise removal through Noise remover Filer by using two filters, One for noise removal and the other to try to improve sounds, reducing the percentage of errors within the audio file.

C Evaluating the dataset: The purpose of the evaluation is to compile feedbacks to feed the system and show how feedback helps learners correct their pronunciation errors. It is also an experience to compile expert assessments of Arabic language teachers. Since the computer does not learn itself on its own unless you give it the key or education mechanism that deals with the values through it or the correct instructions, good results were better, more accurate and more quality than others by standards. Assembling from the last operations previously performed. A program was developed with a personal evaluation interface for each individual expert (where separately evaluator), By running each separate file also not allowed to play the next file until finishing evaluating the current each audio file. That's for all files or audio records, both on its own (evaluating purpose for all audio files), after several teachers or evaluators (expert). And also, to ensure the quality of evaluation, it is preferable to have more than one evaluation, then measure the correlation ratio between them, to reach the end of database step, in general, all of these steps presented in following Figure 3.

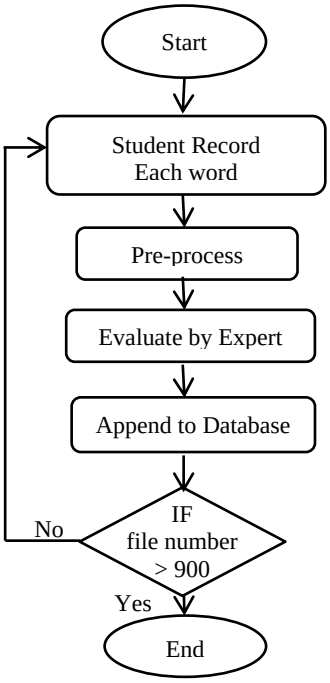

Fig.3. Experimental evaluation using Human experts.

3.2 The proposed method mechanism

After preparing and evaluating the dataset, it became ready to pass to the feature extraction phase to produce a set of feature. These features will be sent to the classification phase to determine if the phoneme pronunciation is correct or not. The following figure illustrates the entire phases of the proposed method. Figure 4 consists of two main phases after loading the dataset; they are feature extraction and classification. So, the following section explains these phases.

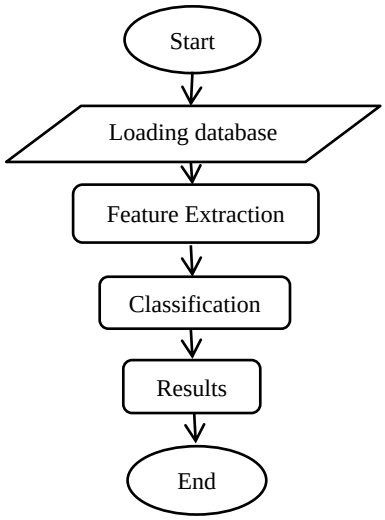

Fig.4. The phases of the proposed system

3.2.1 Feature extraction phase:

Most of the system performance depends on feature extraction step, from this features is fundamental frequency F0 as the main vocal as the lowest frequency of a periodic waveform, also first 4 formants (F1, F2, F3 and F4), occasionally and formant utilized to give acoustic resonance mean of speaker vocal tract. Hence, in phonetics, also can mean either a resonance or the highest spectral output resonance. Formants are often estimated for the sound as amplitude peaks in the frequency spectrum, in the case of the voice. This performs an estimate of resonances to the vocal tract. And MFCC vectors (MFCC0, MFCC1 .... MFCC12) are co-efficients that multilateral. They come from a variety of cepstral that represented for the sound clip (a nonlinear "spectrum-of-a-spectrum"). The frequency bands are spaced evenly on the Mel scale, which closes the human auditory system's reaction more nearly than the linearly-spaced frequency bands handled in the standard cepstrum. That frequency distorted can let better representation of sound.

On the other hand, extracted features can be increased by making the following changes to give new values, and newer traits by previous steps, then instead of limiting the values on average only can use (the mediator, the largest value, the smallest value, the variable value, the mean, and standard deviation). 18 new values, both for the audio file to give different and variable results to name new features of the audio file that can be calculated at the level of each frame within the audio file separately. Assuming that we have [(18 features < F0-F4 (5 types) and 13 values of MFCC > × 6 statistical operations) + The original values of the features to give each file or row for each audio file 126 different values or property].

3.2.2 Classification phase:

For the classification experiments, all material was divided into training (80%) and test data (20%). Furthermore, the material was divided into speech for male and female to developing gender-dependent classifiers. There are two types of phases: Training and Testing. Classification is common in both phases aimed at progress correct utterance identifier. The test pattern is declared to belong to that whose model matches the test pattern best. In training phase, the parameters of the classification model are estimated using the training data. In the testing phase, test speech data is matched with the trained model of each and every class. In this step, some of the classifiers are applied to detect errors for a word from input features and determine if it was uttered correctly or not. There are 900 files or words divided into 10 groups to apply 10 cross-validations to train and test the proposed method. Five classifiers are used, namely, KNN, Support Vector Machine (SVM), Random Forest, Neural Network (MLP), and Na ïve Bayes.

IV. Experiment and Results

The following achievements are to show the best possible results achieved.

4.1. Performance measures:

There are some related measures used in classification include Recall, precision, F-measure and Accuracy.

The recall in (7) is the TP rate (also referred to as sensitivity) what fraction of those that are actually positive were predicted positive? TP / actual positives [26].

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

Precision as (8) is TP / predicted Positive. What fraction of those predicted positive is actually positive? Precision is also referred to as Positive predictive value (PPV) [26];

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

F-measure in (9) is a measure that combines precision and recall which are the harmonic mean of precision and recall [26], the traditional F-measure or balanced F-score:

$$F = 2.\frac{precision.recall}{precision + recall} \tag{9}$$

Other related measures used include True Negative Rate and Accuracy as (10): True Negative Rate is also called Specificity (TN / actual negatives) [26].

1-specificity is x-axis of ROC curve: this is the same as the FP rate (FP / actual negatives) what fraction of those that are actually negative was found to be positive?

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (10)$$

Where: TP (True Positives): quantity of examples predicted positive that is actually positive, FP (False Positives): quantity of examples predicted positive that is actually negative, TN (True Negatives): quantity of examples predicted negative that is actually negative, FN (False Negatives): quantity of examples predicted negative that is actually positive. Classifiers results are shown in Table 2.

4.2. The experiment's result:

This section presents the experimental results based on proposed method. In this work, by using emitting states of various audio files for different students. Following acoustic models are all trained and tested by using WEKA, KNN. The performance is measured based on different parameters, such as experts' Arabic evaluation. Table: 1 shows the average evaluation of the experts using multi averages per experts getting the best average between them, each individual expert evaluates audio files separately by the proposed system. Best average (AVG) of the proposed evaluator system with corrected or true files is (83.8%) and bad or false files (16.2%) (See table 1).

The result show best average evaluation of all experts which entered to system shown in Table 1.

Table 1. The results of the experts' evaluation.

	Expert #1	Expert #2	Total
Best average	159	820	979
Percentage	16.24%	83.75%	100%
Evaluate	0	1	2

After dividing the Arabic letter into groups and prepared as inputs to the system are used to classify the Arabic letters. Also, using two binary evaluation where (0 False forward =>50% and 1 True for other). The system detects errors using KNN algorithm to reach approximately 83.3% percent accuracy. Table 2 shows the results of different classifiers in four measures (accuracy, precision, recall, and f-measure) over the student dataset.

Table 2. The results of the classifier over the dataset.

Classifier	Accuracy	Precision	Recall	F-measure
KNN	83.2821	0.808	0.833	0.816
SVM	84.2051	0.867	0.842	0.771
R.F.	85.0256	0.823	0.85	0.825
N.B	82.1538	0.807	0.822	0.813
MLP	82.359	0.813	0.824	0.818
BAYES	84.4103	0.806	0.844	0.791

V. DISCUSSIONS

The goal of this paper was to create an automatic pronunciation error detection technique and apply it to a speech of a speaker. By investigating the extracted features of the unknown speech and then compare them to the stored extracted features for each different speaker in order to identify correct utterance and detect an error for letters utterance. Also, avoiding as much recognition difficulties and flaws that effect on recognition performance and accuracy. The feature extraction is done by using MFCC, by using functions to calculate the Mel-cepstrum of a signal. In this method, trained models in classification that's prove achieved accuracy, a testing result measure which based on the correlation that was used when matching a perfect evaluation with the speech sound database. But not the perfect accuracy result so, trying KNN classifier to get approximately from ideal recognition accuracy, on the other hand, many experiments using more classifiers with different results of accuracy of recognition. For seeking out to more detection accuracy, noise environment must be prevented as much as possible which considered the main factor for improvement.

Also, it can determine the problem of current research in the presence of deficiencies in the teaching of the Arabic language in terms of exits letters (MAKHARIG) or (Phonemes) and proper pronunciation of the hand. And must not be take consideration individual differences on the other; also, need founded to apply modern technology (ASR) to see how effective they are and use them to enrich the educational process, in terms of helping to fade falling into linguistic errors during the use of the Arabic language to keep pace with the developments of the times.

In addition, finding the lack of studies and Arab research, as well as the lack of adequate database for Arab audios, provide most databases and consider personal efforts. Besides, there are numbers of factors haven't yet been discovered which still remain unsolved

that can reduce the accuracy and performance of error detector, correcting utterance and speech recognition programs where speech recognition process is easy for a human but it is a difficult task for a machine, few factors that are considerable and consider challenge recognizer: homonyms: when the words that are differently spelled and have the different meaning but acquires the same utterance, for example, “there” “their”, “be” “bee”, overlapping speeches: is to understand the speech uttered by different users, current systems have a difficulty to separate simultaneous speeches from multiple users, determining word boundaries: speech are usually continuous in nature and word boundaries are not clearly defined, these happens when the speaker is speaking at a high speed, varying accents: people from different parts of the world pronounce words differently.

This leads to errors in ASR, also plagues human listeners too, large vocabulary items: when the number of words in the database is large, similar sounding words tend to cause a high amount of error i.e. there is a good probability that one word is recognized as the other, noise factor: noise is a major factor in ASR, program requires hearing the words uttered by a human distinctly and clearly, else the machine will confuse and will mix up the words and temporal variance: different speakers speak at different duration. The above mention factors do not provide all support prefect, accurate recognition.

VI. Conclusion

This paper summarized existing research on automated pronunciation error detection as well as some of the work on automated pronunciation error correction.

The remaining challenges that need to be overcome in order to be able to develop truly useful pronunciation teaching applications have also been discussed. Altogether, many components required for such applications already exist. However, the largest remain challenges are availability of a full Arabic database for the pronunciation of all Arabic characters and avoiding noise as much as possible to able to be detected. It has been shown that many different features have been used to measure the various components of pronunciations; there is list of technique with their properties for feature extraction. Through this review, it's found that MFCC are used widely for feature extraction of speech and KNN for model classification containing training and testing is best, in order to achieve a larger degree of accuracy and reliability in Arabic phoneme error detection as expert evaluating to achieve detection of subtle degrees of fluent speech.

References

[1] C. J. Nereveettil, M. Kalamani, and S. Valarmathy, “Feature Selection Algorithm for Automatic Speech Recognition Based on Fuzzy Logic,” pp. 6974–6980, 2014.

[2] M. M. A. Awadalla, “Automatic recognition of Arabic spoken language,” Mansoura University., 2006.

[3] S. D. Shenouda, “Study of an arabic connectionist speech recognition system,” Mansoura University., 2006.

[4] M. Forsberg, “Why is Speech Recognition Difficult?” Technology, pp. 1–10, 2003.

[5] A. M. Ahmad, “Development of an intelligent agent for speech recognition and translation,” 2006.

[6] S. Theodoridis and S. Theodoridis, Introduction to pattern recognition? a MATLAB approach. Academic Press, 2010.

[7] M. F. Abdelaal and EL-Wakdy, “Speech Recognition Using a Wavelet Transform,” 2008.

[8] D. MANDALIA and P. GARETA, “Speaker Recognition Using MFCC and Vector Quantization Model,” Electronics, vol. Program, C, no. May, p. 75, 2011.

[9] M. Al Hawamdeh, “Loud Reading Errors of Third Grade Students in Irbid Governorate and their Relationship to Some Variables,” 2010.

[10] Ahmed, Abdelrahman, Yasser Hifny, Khaled Shaalan, and Sergio Toral. "Lexicon Free Arabic Speech Recognition Recipe." In International Conference on Advanced Intelligent Systems and Informatics, pp. 147-159. Springer International Publishing, 2016.

[11] Ewees, A. A., Mohamed Eisa, and M. M. Refaat. "Comparison of cosine similarity and k-NN for automated essays scoring." cognitive processing 3, no. 12 (2014).

[12] Reafat, M. M., A. A. Ewees, M. M. Eisa, and A. Ab Sallam. "Automated assessment of students arabic free-text answers." Int J Cooperative Inform Syst 12 (2012): 213-222.

[13] Menacer, Mohamed, Odile Mella, Dominique Fohr, Denis Jouvet, David Langlois, and Kamel Smaili. "An enhanced automatic speech recognition system for Arabic." In The third Arabic Natural Language Processing Workshop-EACL 2017. 2017.

[14] A. E.-R. S. A. El-Rahman, “Computer Aided Pronunciation Learning for Arabic Language,” 2007.

[15] T. A. F. I. Sheisha, “Building A speech recognition system for spoken Arabic,” Cairo University - Institute of Statistical Studies and Research, 2009.

[16] H. S. A. Abdelaziz, “Language speech impairment rehabilitation using automatic speech recognition (ASR) technique,” Cairo University - Faculty of Engineering, 2013.

[17] E. M. M. Essa, “Arabic speech recognition,” 2008.

[18] N. A.-S. B. S. Ahmed, “Distributed Speech Recognition of Arabic Speech over GSM Channels,” Cairo University., 2010.

[19] S. Mohanty and B. K. Swain, “Speaker Identification using SVM during Oriya Speech Recognition,” Int. J. Image, Graph. Signal Process., vol. 7, pp. 28–36, 2015.

[20] A. Pahwa and G. Aggarwal, “Speech Feature Extraction for Gender Recognition,” Int. J. Image, Graph. Signal Process., vol. 8, pp. 17–25, 2016.

[21] M. Ahmed, P. C. Shill, K. Islam, and M. A. H. Akhand, “Acoustic Modeling of Bangla Words using Deep Belief Network,” Int. J. Image, Graph. Signal Process., vol. 7, pp. 19–27, 2015.

[22] Vimala. C. and V. Radha, “Efficient Acoustic Front-End Processing for Tamil Speech Recognition using Modified GFCC Features,” *Int. J. Image, Graph. Signal Process.*, vol. 8, pp. 22–31, 2016.

[23] M. K. noby Khalil, “Development of a Cognitive Speech Recognition System to Improve the Pre-school Children speech Abilities a thesis,” 2011.

[24] M. A. Imtiaz and G. Raja, “Isolated word Automatic Speech Recognition (ASR) System using MFCC, DTW & KNN,” in 2016 Asia Pacific Conference on Multimedia and Broadcasting (APMediaCast), 2016, pp.

106–110.
[25] H. Y. F. M. Elghamrawy, “Improving Arabic phonemes recognition using nonlinear features,” Cairo University., 2013.
[26] Powers, D.M., Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation, 2011.

Authors’ Profiles

Moner N. M. Arafa received the B.Sc. degree in Preparing Computer Teacher Department, Faculty of Specific Education, from Mansoura University, Egypt in 2011. His research interest includes pattern recognition, speech processing, and advanced machine learning.

Reda Elbarougy received his B.Sc., and M.Sc., degrees from Mansoura University, Egypt, in May 1997, and February 2006, respectively. Both were in computer science. He was with the Faculty of Science, Mansoura University from 1999 to 2009. In July 2009, he joined the Japan Advanced Institute of Science and Technology (JAIST), Japan, as a Ph.D. student. Since 2014, he has been an Assistant Professor with Mathematics Department, Faculty of Science, Damietta University. Currently he is a post-doctor researcher funded from JSPS to conduct a research in Japan Advanced Institute of Science and Technology (JAIST) from June 2017 till now. His current research interests include speech analysis, speech emotion recognition, and synthesis.

A. A. Ewees work as assistant professor at Computer department, Faculty of Specific Education, Damietta University, Egypt. He co-advises master and Ph.D. students, as well as leading and supervising various graduation projects. His research interests include machine learning, pattern recognition, natural language processing, computational intelligence, and artificial intelligence.

G. M. Behery received the B. Sc. degree in computer science from the faculty of science, Suez Canal University, Egypt, in 1984, and the M.Sc. degree in computer science from Mansoura University, Egypt, in 1989 and the Ph.D. degree in computer science from Mansoura University (Egypt)/Friedrich-Alexander University (Erlangen-Nürnberg - Germany), in 1993. From 1993 to 2008, he was Assistant professor with Mansoura University. In 2008, he has been an Associate Professor in computer science with Mansoura University. In 2016, he has been a Professor in computer science at Damietta University. His current research interests include image processing, pattern recognition, animal recognition, AI, Neural networks and computer language design.

How to cite this paper: Moner N. M. Arafa, Reda Elbarougy, A. A. Ewees, G. M. Behery," A Dataset for Speech Recognition to Support Arabic Phoneme Pronunciation", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.10, No.4, pp. 31-38, 2018.DOI: 10.5815/ijigsp.2018.04.04

