Recognition of Correct Pronunciation for Arabic Letters Using Artificial Neural Networks

Abeer M. K. Osman[1], Hussain A. Ibrahim[2, 3], Mohamed Adany[4]

[1]Faculty of Computer Sciences and Information Technology, Sudan University of Sciences and Technology (SUST), Sudan
[2]Biomedical Engineering Dept., Sudan International University, Sudan
[3]Biomedical Engineering Dept., Gezira College of Technology, Sudan
[4]Computer Sciences Dept., Blue Nile University, Sudan
Abeer.mohammedkheir@gmail.com

Received: 21/09/2019
Accepted: 31/10/2019

Abstract- Automatic speech recognition (ASR) plays an important role in taking technology to the people. There are numerous applications of speech recognition such as direct voice input in aircraft, data entry and speech-to-text processing. The aim of this paper was to develop a voice-learning model for correct Arabic letter pronunciation using machine learning algorithms. The system was designed and implemented through three different phases: signal preprocessing, feature extraction and feature classification. MATLAB platform was used for feature extraction of voice using Mel Frequency Cestrum Coefficients (MFCC). Matrix of MFCC features was applied to back propagation neural networks for Arabic letter features classification. The overall accuracy obtained from this classification was 65% with an error of 35% for one consonant letter, 87% accuracy and an error of 13% for 10 isolated different letters and 6 vowels each and finally 95% accuracy and an error of 5% for 66 different examples of one letter (vowels, words and sentences) stored in one voice file.

Keywords: MFCC features, neural networks, classification.

المستخلص- تؤدي أنظمة التعرف على الأصوات الرقمية دورا هاما في تعليم اللغات المنطوقة وتعليم الحروف. هناك الكثير من التطبيقات المتنوعة لأنظمة التعرف على الصوت بإستخدام جهاز الحاسوب مثل الطائرات وأنظمة البيانات وعمليات تحويل النص الى صوت. الهدف من هذا البحث هو تطوير نظام صوتي لتعليم الحروف العربية للأطفال بإستخدام تعلم الآلة. يمكن تقسيم أنظمة التعرف على الصوت الى ثلاث مراحل رئيسية هي: معالجة الإشارة الرقمية واستخراج خصائص الصوت وتصنيف الصوت. تم استخدام منصة الماتلاب كأداة هامة لإستخراج خصائص الصوت المعتمدة بناءاًعلى تقنية (معاملات سبسترم لترددات ميل). تم تدريب الشبكة العصبية على هذه المعاملات المستخرجة من الاشارة الصوتية الداخلة حيث تم حساب دقة التصنيف النهائي لنطق الحرف لتصل الى 65% وبلغت نسبة الخطأ الكلي للتصنيف 35% لحرف واحد ساكن وبلغت دقة التصنيف النهائي 87% ونسبة خطأ 13% لعشرة أحرف مختلفة بست حركات لكل منها بينما بلغت دقة التصنيف 95% ونسبة خطأ 5% عندما استخدم 66 مثال مختلف لحرف واحد من حركات وكلمات وجمل تم تسجيلهم في ملف صوتي واحد.

INTRODUCTION

Arabic language is the fifth language in terms of the number of speakers[1]. All Muslims when recite verses of the holy Quran in praying or in other situations, they should follow the rules of pronunciation. Little researches on Arabic speech recognition have been applied compared to other languages (e.g. Spanish or Mandarin) [1].

In Sudan, people have common and major difficulty with pronouncing some Arabic letters such as: (“ظ” “ض”) and (“ت” “ط”), because they have relative acoustic outputs and features,

especially primary students. There is high motivation to design an accurate recognition system for primary levels, Quranic schools (Khalawi) and non-native speakers to get efficient and clear pronouncing of Arabic letters.

Speech recognition technology is one of the recent computer technologies. Mel-Frequency Cepstral Coefficients (MFCC's) is the relationship between Human ear's critical bandwidths with frequency. It is used for analyzing and extraction of pitch vectors [2]. Machine learning and pattern recognition classify data into different groups, fit curves or make predictions.

Problems such as speech recognition and computer vision are too complex to solve analytically. Computer algorithms can help to solve them iteratively[3]. Artificial neural networks (ANN) is subfield of pattern recognition lying on mimicking biological neurons with simple neural networks using electrical circuits[3]. Two different ways to use neural networks for acoustic modeling, namely prediction and classification of the speech patterns[4].

Recognized three spoken Arabic letters of hijaiyah is presented in [4] which having the same pronunciation by Indonesian speakers but has different makhraj (place of the letter out) in Arabic letters of sa, sya and tsa. (MFCC) was used for feature extraction and (ANN) were used for classification. The average accuracy was 92.42%, and each letters (sa, sya, and tsa) has accuracy of 92.38%, 93.26% and 91.63% respectively[4]. [5] developed Quran reciter recognition system based on (MFCC) feature extraction and artificial neural network (ANN) model. A database of five Quran reciters created, trained and tested and accuracy of classification was 91.2% [5].

Samiksha Sharma, Anupam Shukla and Pankaj Mishra [6], recognized speech four different languages and words Hindi, English, Sanskrit and Telugu. MFCC and delta-MFCCs were applied for feature extraction and ANN was used to classify the input as a set of 18 words and languages. In first experiment, the overall sound recognizer's performance was 83.89% and language recognizer's performance was 83.3%.

In second experiment, radial basis function network used as classifier and overall word recognition accuracy was 91.7% and language recognition accuracy was 91.1% [6].

The main aim of this paper was to develop an Arabic letters recognition system that can record voice signal, recognize its features using (MFCC) techniques and classify the final output as being the correct the spoken/pronounced letter or not using artificial neural networks.

MATERIALS AND METHODS

The implemented model consists of the following stages:

Digital signal pre-processing:

Reading and extracting the voice data:

Arabic letters database has been installed from (Lebanese Arabic from Scratch - Alphabet Book). It contains 10 different alphabetic Arabic letters, 6 vowels for each, with high quality voice recorded, each of which with sampling frequency of 44000 Hz and voice clip duration of 2 seconds. The total size is 60 different files, so the overall examples for training the neural network are 3600 samples of frames (60 frames each). The letters used are: Ain (ع), Gin (غ), Taa (ت), Tta (ط), Haa (ح), Kha (خ), Sud (ص), Daa (ض), Dha (ظ) and Tha (ذ).

Segmentation:

To process analog signal digitally, it has to be converted first into digital form. The segmentation of speech depends on two methods, manual and automatic. In this work, auto segmentation was chosen. All sounds do not start from sample 1, although most of them are posted in the beginning of the carrier. The reason is that there was silence in the beginning of each sound. Each voice clip contains two parts: voiced part, unvoiced part and the silent part.

There are many techniques used for silence removal and one of these applied here was Short Term Energy method (STE). It was calculated for all the frames of the original signal. STE threshold value was determined and labeled for all minimum values of amplitudes in all frames. The regions of silence were extracted and removed from both sides of voice carrier.

Frame blocking, windowing, filtering and DFT analysis:

Frame blocking:

Speech signal is a continuous signal and it was divided into frames for proper analysis. The original speech signal (after silence removal) was named here as SR. It was divided into frames; each one has a duration of 0.02 seconds (20 msec).

Windowing and filtering:

There are many types of window functions used in digital FIR filtering (Finite Impulse Response) and one of these selected here was Hamming window which was multiplied by the chosen frame and then filtered it using high pass filter to specify only the useful (voiced) information block and isolate the low frequencies part which contains no useful information. Hamming window was designed in MATLAB with a size like each frame size.

Discrete Fourier Transform (DFT) analysis:

The computed Discrete Fourier Transform (DFT) analyzes the frequency components of processed signal.

1. Feature extraction:

Feature extraction is the process of extracting the fundamental parameters identifying a speech signal. The linear frequency scale was converted into Mel scale and the maximum number (M) of equal spaced frequency filter triangular banks was determined to be 39 filter bands, starting from zero frequency and ending with a maximum frequency of approximately 500 Hz.

Because there were a problem and difficulty in equalizing the sampling frequency, frame size, frame duration and hence maximum number of frames between both the target letter voice data and the actual spoken data, so this was solved by getting the filter banks for all the frames in the voice signal. No. of periodograms (Normalized power) were also computed to get the triangular shaped filter banks.

MFCC matrix was generated and so for easy analysis and classification later, it was converted into a column vector of 39x1 size. Finally, the logarithm of this function was also computed to get the equivalent cepstral coefficients.

2. Classification:

In this paper, artificial backpropagation neural networks have designed, trained, validated and tested for three experiments.

First Experiment

Only the neural network was used for training one letter (consonant) without vowel. Backpropagation neural networks was designed for this goal with one input layer (15 neurons equal to the number of cepstral coefficients), one hidden layer (100 elements) and one output layer (one node for one of either two numerical classes 0 standing for not the target latter and 1 standing for correct letter).

Second Experiment

The neural networks were used for training 10 letters with six vowels each. The total is 60 different vowels.

Backpropagation neural network was designed for this goal with one input layer (39 neurons equal to the number of cepstral coefficients), one hidden layer (48 elements) and one output layer (60 nodes for 60 classes of different vowels of 10 letters).

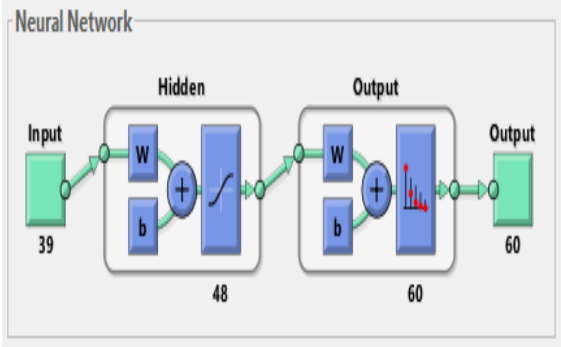

Figure 1. The designed neural network for experiment two

The target data was selected here from the same MFCC features matrix but converted first into binary (logic) values (multiplied each element in matrix with -1 to eliminate the negative small decimal numbers for proper classification purpose) and then compared with the same train data input to the network.

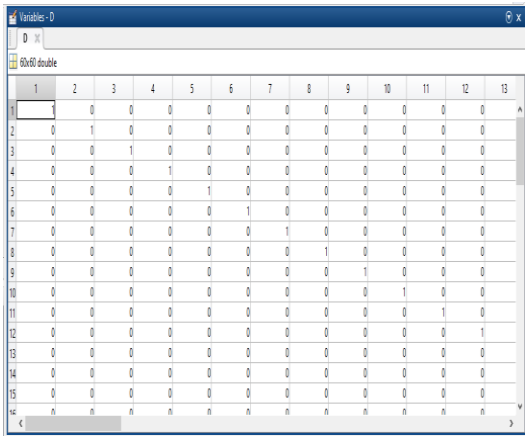

Figure 2. Created target data for classification of ten letters database

The total number learning epochs was 8000 times, the error goal Mean Square Error (MSE) was used) was to be 1e-25 and the learning rate was 0.01. There was a

problem in accurately adjusting these training parameters to get high level of performance.

Therefore, the solution was in increasing up the number of input (spoken cepstral coefficients) but not increasing the number of hidden layers because of more training time delay might been occur.

The suggested number was 39 of cepstral coefficients and by default MATLAB considers about 60% of the input data for training, 20% for validation, i.e. to support the training and finally 20% for testing. The goal was in increasing the size training data to ensure proper performance of fitting the goal line of classification (Y=T), where Y is the testing data and T is the target data.

The 39 cepstral coefficients of the 10 different spoken and recorded letters (6 vowels each), were applied to the input layer (39 neurons), trained, validated and then finally tested against the target cepstral coefficients gotten from the feature extraction process of the standard letters. All training data frames examples has been resized to be of size (39x60) to unify them with the target data size.

Although there is no specific rule or concept to create the target data, i.e. just depending on the type of research problem or objective in other words, so here it was designed for the second experiment to be either 0 or 1, i.e. 1 to indicate the classified (target) letter and 0 to indicate nothing, i.e. not the target one.

For example, in Figure 2, the first column indicate the target letter (Ain) (ع) with vowel Fatha, so the 1 in the 1st pixel is opposite immediately to the 1st output node and the rest of the rows in the 1st column are 0 to indicate that no other letter to be classified in this class.

Third experiment

The neural network was used for training 66 examples of one letter e.g. Ain (ع) all loaded in one voice file (.wav).

The letter is spoken at first as 6 vowels, uttered one time with one letter, two letters, three letters, in complete word and finally in a sentence as shown in Figure 3. It contains higher number of samples and frames (957 frames) and this is useful for achieving higher accuracy of classification.

The target data was designed as 2x957 size, i.e. two classes for each frame. Each 15 frames represent an example in file.

RESULTS AND DISCUSSION

Voice signal has 2 sec recorded clips. In Figure 3, the signal contains about 100,000 samples.

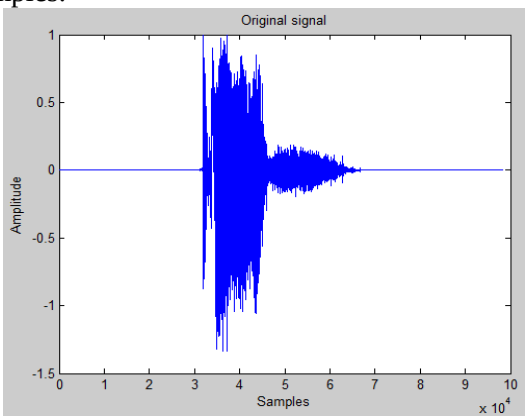

Figure 3: The original voice signal data

There was delay in the signal. The silence part on either side is very clear because the voice was recorded using very high-quality microphone and in advanced studios. or proper scaling and silence removal, STE was normalized. The frame energy (STE) was displayed together as red step lines as shown in Figures 4 to determine the threshold accurately of cutting the silence/noisy parts.

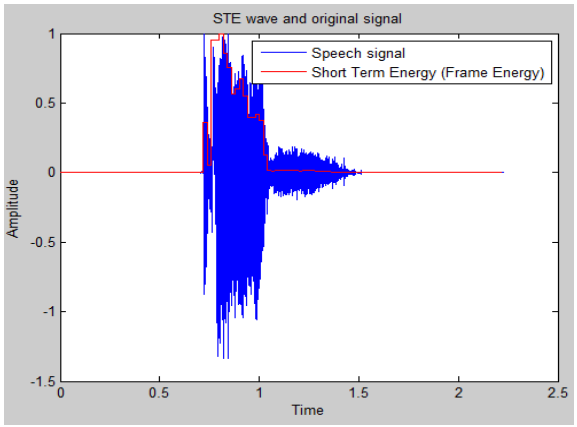

Figure 4: STE wave with original voice data

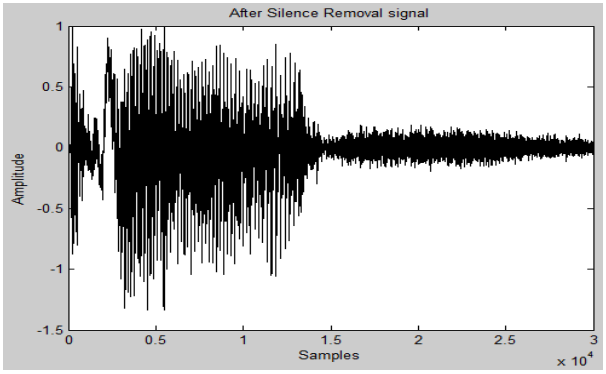

Figure 5: The voice data after silence removal

The result of silence removal from Figure 4 was very efficient. Note that the total number of samples in Figure 3 was reduced to less than half as shown in Figure 5. Frames in Figure 6 were selected to be at position of 15, the frame size of each has 882 samples and both look like periodic waveform. After silence removal, the high frequency band was increased as shown in Figure 7 after FFT was applied.

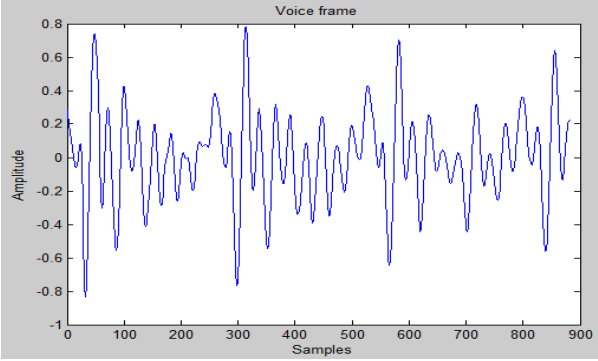

Figure 6: The frames of the original data at position 15

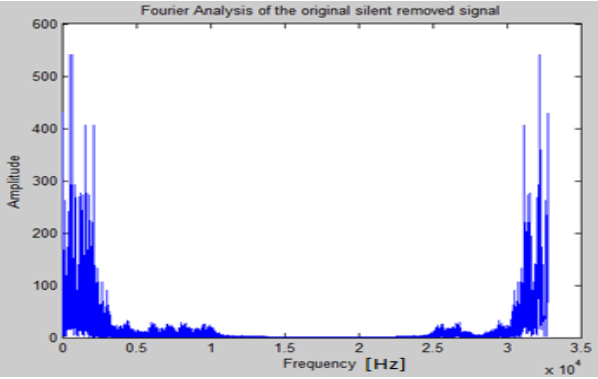

Figure 7: FFT of the original silent removed signal

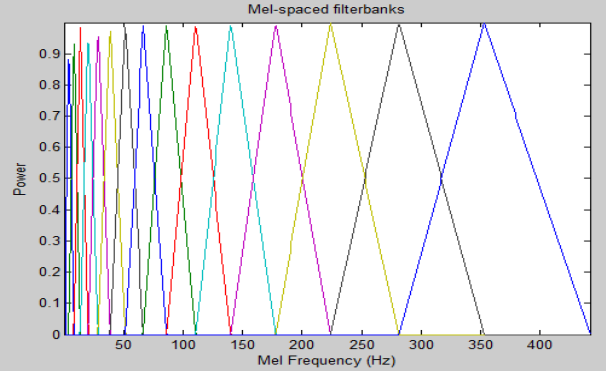

Figure 8: Mel spaced triangular filter banks

In Figure 8, the filter banks were increased in their bands as they move to the right which means more useful information at high frequency elements.

Figure 9 demonstrates the log power scale of MFCC confidents in each filter bank for all the frames and the same mechanism can be applied to the rest of all letters. After the frequency 30 Hz, most of the curves tend to be at steady state between the range (-5 dB – (-10) dB) but before it there was fluctuations in dB power around 0 dB and (-5) dB.

It showed the inverse relationship between the frequency banks and power, i.e. with increasing the frequency the power in dB will decreased gradually till reaching the most possible lower values, i.e. (-10 dB) and then tend to stabilize.

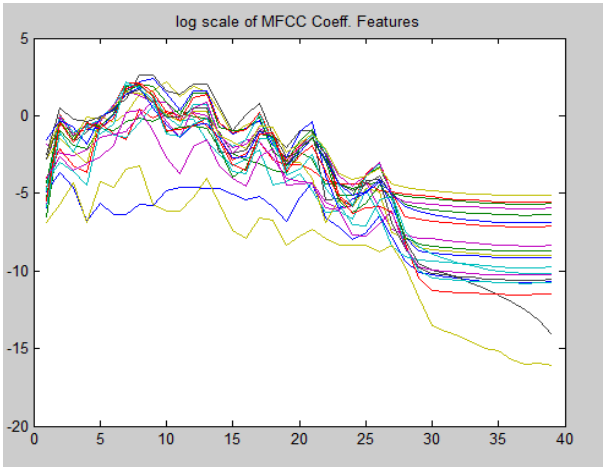

Figure 9: Log scale of MFCC coefficients features of voice signal

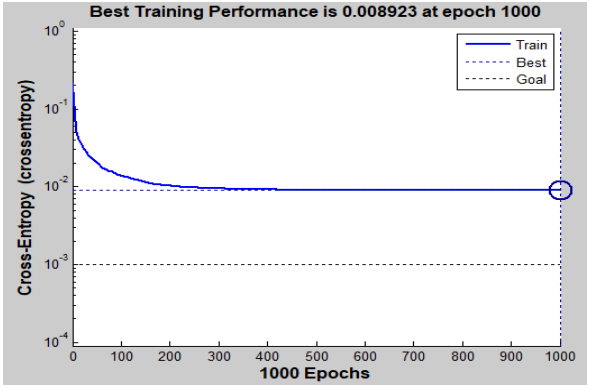

Figure 10: Best training performance plot of MFCC coeff.

The neural network was trained 2000 times of epochs with a goal reached 1e-9. The actual goal of performance was achieved 0.0089 at epoch

1000. The network needs to be trained with more times of iterations to achieve the maximum degree of accuracy.

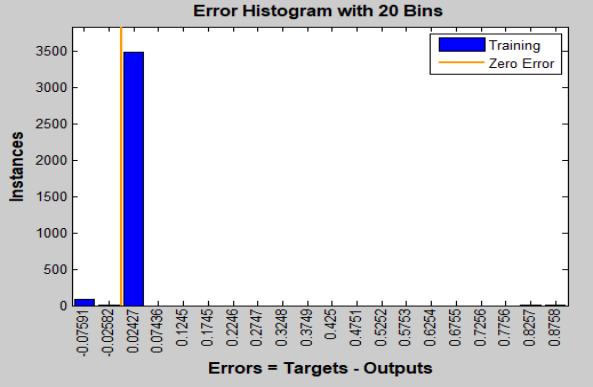

Figure 11: Error histogram of trained neural networks

TABLE 1: COMPARISON OF BOTH OVERALL ACCURACY AND ERROR BETWEEN THE THREE EXPERIMENTS

No. of experiment	Overall accuracy achieved	Error
Experiment one	65%	35%
Experiment two	87%	13%
Experiment three	95%	5%

In Figure 11, the variation of errors was distributed semi-normally which stands for enhanced design and proper MFCC features extraction was obtained from this system. From Table 1, the highest overall accuracy was achieved for experiment three because it has higher amount of input data examples although the semi-adjustment design procedures with other experiment one and two but the last both have been designed for less number of input data.

The experiment one was the lowest accuracy because it has only conducted for one consonant letter, so it was difficult to recognize it properly.

CONCLUSION AND RECOMMENDATIONS

Phenome recognition system consists of three different phases: Digital signal processing, voice features extraction and neural networks for classification. Mel frequency filter banks of 39 non-linearly spaced, when applied to the spectral magnitudes of FFT yields the dominant frequency components (or peaks or formants) for each frame. MFCC algorithm performed on 'wav' files in MATLAB yields a matrix with number of columns equal to number of frames, which was determined by the size of input files and number of row equal to the DCT size, which was 39 in our case for voice signal. The back propagation multilayer neural networks have designed and tested among three different experiments and trained much of times to approach higher degree of accuracy.

The overall accuracy obtained from this classification was 65% with an error of 35% for one consonant letter, 87% accuracy and an error of 13% for 10 isolated different letters and 6 vowels each and finally 95% accuracy and an error of 5% for 66 different examples of one letter (vowels, words and sentences).

In future, Hidden Markov Model (HMM) should be developed to enhance the accuracy of phenome recognition, also coding with Python language is the most efficient, applicable and simple programming language for machine learning and finally this system should be developed for learning Tajweed and Tellawah of Qur'an.

REFERENCES

[1] R. M. E. A. Moaz Abdulfattah Ahmad, (2011), "Phonetic Recognition of Arabic Alphabet letters using Neural Networks," Int. J. Electr. Comput. Sci., vol. 11, no. 1.

[2] S. O. M. Nssr, (2016), "Voice Recognition by using Machine Learning A Case Study of some Rules of Tajweed," Sudan University of Science and Technology College.

[3] Staven, (2016), "Detection of phonetic features for automatic classification of Norwegian Dialects," Norwegian University of Science and Technology.

[4] E. S. Wahyuni, (2017), "Arabic Speech Recognition Using MFCC Feature Extraction and ANN Classification," in 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering.

[5] H. M. Tayseer Mohammed Hasan Asda, Teddy Surya Gunawan, Mira Kartiwi, (2016), "Development of Quran Reciter Identification System Using MFCC and Neural Network," TELKOMNIKA Indones. J. Electr. Eng., vol. 17, no. 1, pp. 168–175.

[6] F. B. T. Hassan M. H. Mustafa, (2016), "On Comparative Study for Two Diversified Educational Methodologies Associated with "How to Teach Children Reading Arabic Language? Neural Networks' Approach," Open Access Libr. J., vol. 3, no. e3186.