Journal of Engineering Science and Technology
Vol. 16, No. 6 (2021) 4853 - 4863
© School of Engineering, Taylor's University

A NEW SYSTEM FOR CONVERTING VOICES OF PERSIAN LETTERS INTO GESTURE

SHAKER K. ALI[1,*], ZAHOOR M. AYDAM[1], WAMIDH K. MUTLAG[2]

[1]Computer Sciences and Mathematics College, University of Thi_Qar, Thi_Qar, Iraq
[2]Al Shatrah Technical Institute, Southern Technical University, Iraq
*Corresponding Author: shaker@utq.edu.iq

Abstract

Blind people cannot communicate with dumb people in light of the fact that the visually impaired can't see the signal , the visually impaired individuals can just stand up them thoughts , while the moronic individuals can just use (visual) the motions gestures, so on the off chance that the visually impaired individuals utilize the voices can be changed over into motions, the communication between daze blind people and dumb peoples is easy. In this paper, communication and understanding the proposed system to change over the voices of Persian letters into signals (pictures relating to Persian letters) the recommended system gathered voices of Persian letters from various people (five people). This information is basely containing five voices for each letter (160 voices altogether) since there are 32 letters in Persian language. The proposed system is divided into two sections; the initial segment is for preparing, and the subsequent second part is for testing .The features separated relying upon MFCC and group classified by Linear Discriminate Analysis (LDA) and Quadratic Discriminate Analysis (QDA). The final results accuracy of suggested system is 96.875%.

Keywords: Gesture and voice recognition, Linear Discriminate Analysis (LDA), MFCC, Persian letters, Quadratic Discriminate Analysis (QDA).

1. Introduction

Most research on gesture focus in how to analyse the gesture whether the gesture is letters, words or numbers, and extract a few highlight features classify gestures; a few studies convert these features into voices; in this paper, the reverse done.

Some studies provide an electronic reading system to text over to voice utilized by outwardly disabled and visually impaired individuals with an inability to read and relates in particular to the electronic reading system of Persian letters [1].

The most extreme significant part in voice acknowledgment is the amount to separate the highlighting extraction features of voices; it needs explicit calculations that may apply in the feature extraction cycle and classification process . There are a lot of feature extraction algorithm like as Linear Predictive Coding, Mel Frequency Cepstral Coefficients and Linear Predictive Cepstral Coefficients, etc. So, you need an appropriate algorithm to be used in the proposed system [2].

The main algorithms in discourse signal speech examination are Linear Predictive Coding, Mel Frequency Cepstral Coefficients and Linear Predictive Cepstral Coefficients [3]. Notwithstanding, the MFCC technique determination in the recommended system since it contains it has an ideal precision level on the speech recognition system [4].

After extracting the features, the following stage is to classify into group the voices , likewise many algorithms for voices order like LDA and QDA these algorithms had been used especially in the example pattern recognition system since it is exceptionally useful as the technique for information compression . LDA changes noticed variable to be one collection of linear collection that don't relate, called the aster component. The all out of the master component is not exactly or same as the absolute variable [5]. By using of gather MFCC and (LDA, QDA), it was required to increase the accuracy of the proposed system and limit the feature data distance.

2. Proposed System

The proposed system contains two sections, one for preparing, and the next part is for trying. The preparation part comprises four stages; the initial step is data(voice) assortment collection, which gathered representatives of 32 letters from five people. The subsequent advance is for examination and concentrates highlight extract features by utilizing MFCC. The third step is for voice letters arrangement by using LDA and QDA ,and the fourth step is for changing over each letter type into a picture (motion) for dumb people; the proposed system stream chart is as demonstrated in Appendix B.

2.1. Database collection (voices collection)

There is no database available (online) that to use as a reference for other researchers, only some learning videos for learning words or numbers or letters. So, dataset collected from 5 different persons (males and females) by recording voices from 5 different persons; each, person recorded 32 voices for every 32 letters, so the total dataset is 160 voices (160 voices for every 32 letters), which have been recorded in same environment. All voice signals have under various conditions, like the record time length, and sound amplitude level. The recorded voices are stored in format “.WAV" extension. Postulates to user for choosing any sample of the speech for testing from recorded dataset

2.2. MFCC Features

This technique is a widespread common technique in extracting the distinctive features of sound in speech recognition systems due to the accuracy of its results and the ability to partially eliminate the noise of the signal [6] As well as the speed of application and less complex and more effective under different circumstances In this technique, the human hearing process is approached, i.e., an attempt to extract the signal characteristics in a manner consistent with the human hearing mechanism, since the human ear is sensitive to frequencies that are less than 1000 Hz and weak to frequencies higher than 1000 Hz. The reason for the design of such filters is that the human ear is not sensitive to high frequencies and therefore can reduce the number of filters characteristic of these frequencies . It has been discovered that MFCC is commonly utilized for extracting speech features due to its robustness to noise [7]. MFCC is used due to the following reasons . In the MFCC, the voice signal passes through the next stages, as demonstrated in Fig. 1:

1. MFCC is the most significant feature that are needed between various speech application types.
2. Provides highly accurate results for cleaning speech.
3. MFCC may be considered a standard feature in speech recognition systems.

Step1: Pre-emphasis filtering

This filter has been applied to voice signal to increase the high-frequency energy and reduce the low–frequency energy, as defended in Eq. (1) [8].

$$q_t = \alpha * q_t * (1 - \alpha) * q_{(t-1)} \tag{1}$$

where: α is a pre-emphasis filter constant, and the value usually set as $0.9 < \alpha < 1.0$. q_t is a frequency energy.

Step 2: Frame blocking

Divided the voice signal into various number of edges, where each frame edge comprises of *M*-sample examples and these edges are contiguous and separated from one another by *N*-tests. If (*N* <=*M*)

The contiguous edge frames cover, from this standard the acoustic sign divided into several overlapping frames by *N*-test samples [9].

Step3: Windowing

Each casing went through a window for a specific period to reduce the discontinuity of the signal from the beginning and end of each frame. The Eq. (2) characterize the window *W*(*m*) and input to signal *X*(*m*) at that point yield signal is $Z(m)$[10].

$$Z\,(m) = X(m)W(m), 0 \le m \le M- 1 \tag{2}$$

where *W*(*m*) is the window function, $0 \le m \le M - 1$, and *N* is the quantity number of samples in each frame. The most normal window in speech recognition systems is the Hamming window and knowledge information of the equation. In this paper, the Hamming window utilized, as represented in Eq. (3) [7]:

$$W(m) = 0.54 - 0.46\, cos\left(\frac{2\pi m}{N} - 1\right), 0 \leq m \leq M - 1 \tag{3}$$

Step 4: Fast Fourier transform (FFT)

Is a quick application of Discrete Fourier transform which converts M frames into frequency domain, the Eq. (4) explain FFT [10].

$$\mathrm{Y}\,(\mathrm{u}, \mathrm{v}) \sum_{x=0}^{N-1} \sum_{\mathrm{y}=0}^{\mathrm{M}-1} \mathrm{X}(\mathrm{x}, \mathrm{y}) \mathrm{e}^{-\mathrm{j}2\pi\left(\frac{\mathrm{ux}}{\mathrm{N}}+\frac{vy}{M}\right)} \tag{4}$$

where : $X(x,y)$:is input signal x=0,1,2,. . ., N-1 and y = 0,1,2,. . ., M-1 denote a digital image of size N*M pixels , u=0,1,2, . . ., N-1 and v=0,1,2, . . ., M-1.

Step5: Mel filter bank

It is composed of overlapping triangular filters with cut frequencies determined by the central frequencies of the neighbouring filters. Filters have linear spaced frequencies and a fixed frequency range on the Mel scale. as shown in Eq. (5) [6, 11].

$$F(Mel) = 2595 * log\ 10\ (1 + \left(\frac{F}{700}\right) \tag{5}$$

where $F(Mel)$ is mel filter bank.

Step6: Discrete cosine transform (DCT)

To re-signal ,into time domain form in this step has done by using a Discrete Cosine conversion .DCT transforms the cosine component only. DCT 2-D for image $X(x,y)$ with length N*M can be expressed in Eq. (6) [8].

$$C(u,v) = \frac{2}{\sqrt{N*M}} \alpha(u)\alpha(v)X(x,y) cos\left(\frac{\pi(2x+1)u}{2N}\right) cos\left(\frac{\pi(2x+1)v}{2M}\right)$$

For u=0,1,2, . . ., N-1 and v=0,1,2, . . ., M-1.

Step7: Cepstral coefficient

The end results are (13) Cepstral coefficients that are used a voice recognition feature.

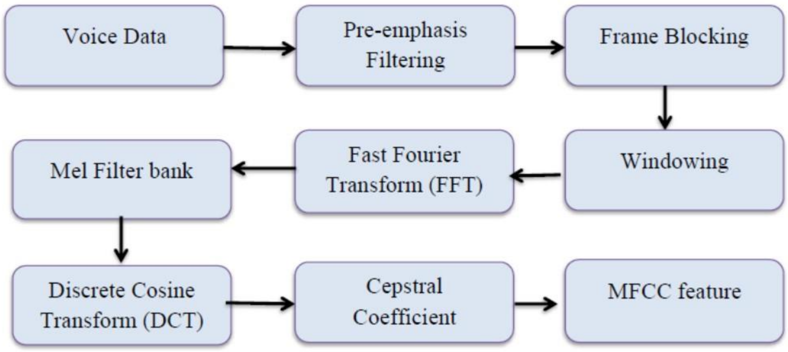


Fig. 1. Block diagram of MFCC feature extraction [12].

3. Discrimination Analysis Algorithm

Discrimination analysis is one method of classification. It assumes that the different categories of data generated based on different Gaussian distributions [13]. Two types used in proposed system; Linear Discriminate Analysis (LDA) and Quadratic Discriminate Analysis (QDA) Both LDA and QDA can be derive from simple probabilistic models that model the class conditional distribution of the data $P(Y\backslash x = k)$ for each class. Probability can also obtain by using Bayes' rule as shown in Eq. (7):

$$P\left(x = \frac{k}{Y}\right) = \frac{P(Y\backslash x=k)P(x=k)}{P(Y)} = \frac{P(Y\backslash x=k)P(x=k)}{\sum_l P(Y\backslash x=l)P(x=l)} \tag{7}$$

We select the class k which maximizes this conditional probability. More specifically, for linear and quadratic discriminant analysis, $P(Y, x)$ is modeled as a multivariate Gaussian distribution with density as shown in Eq. (8) [7]:

$$P(Y/x = k) = \frac{1}{(2\pi)^{\frac{d}{2}}|\sum_{k\cdot}|^{\frac{1}{2}}} exp\left(-\frac{1}{2}(Y - M_k)^t \textstyle\sum_k^{-1}(Y - M_k)\right) \tag{8}$$

where d is the number of features, M_k is mean vector of each class and Y is data vector. To use this model as a classifier, we just need to estimate the class priors $P(x = k)$ from the training data (by the proportion of instances of class K), the class means M_k (by the empirical sample class means) and the covariance matrices (either by the empirical class covariance matrices, or by a regularized estimator). In the case of LDA, the Gaussians for each class are assumed to share the same covariance matrix: $\sum_K = \Sigma$ for all k.

This leads to linear decision surfaces, which can be clear by comparing the log-probability ratios as shown in Eq. (9), $log\ [P(x = k|Y)/P(x = l|Y)]$

$$log\left(\frac{P(x=k\backslash X)}{P(x=l\backslash Y)}\right) = log\left(\frac{P(Y\backslash x=k)P(x=k)}{P(Y\backslash x=l)P(x=l)}\right) = 0 \leftrightarrow$$
$$(M_k - M_l)^t \textstyle\sum^{-1} Y \ = 1/2(M_k^t{\_}M_l^t \textstyle\sum^{-1} M_l) - log\frac{P(x=k)}{P(x=l)} \tag{9}$$

In the case of QDA, there are no assumptions on the covariance matrices of the Gaussians, leading to quadratic decision surface [14].

4. Convert Features into Gestures (Images)

The outcome resultant from the last step is features vectors of the voices for Persian letters type

(أ, ب, پ , ت, ث, ج, چ, ح , خ, د, ذ ,ر, ز, ژ , س, ش, ص, ض, ط, ظ
ع, غ, ف, ق, کـ, گ, ل, م, ن, ه, و ,ي)

So, it will be converted into appropriate image depend on the letter , where each letter has appropriate gesture (image). As shown in Fig. 2.

5. Experimental and Results

The proposed system utilizes 160 voice letters, which divides into two groups 1. 128 for training (preparing) and 32 for testing. The highlighting features extracted from each letter voice will be classified by using LDA and QDA, the Figs. 3(a), (b), and (c) show the highlight features extracted from voice letter (أ) in the training

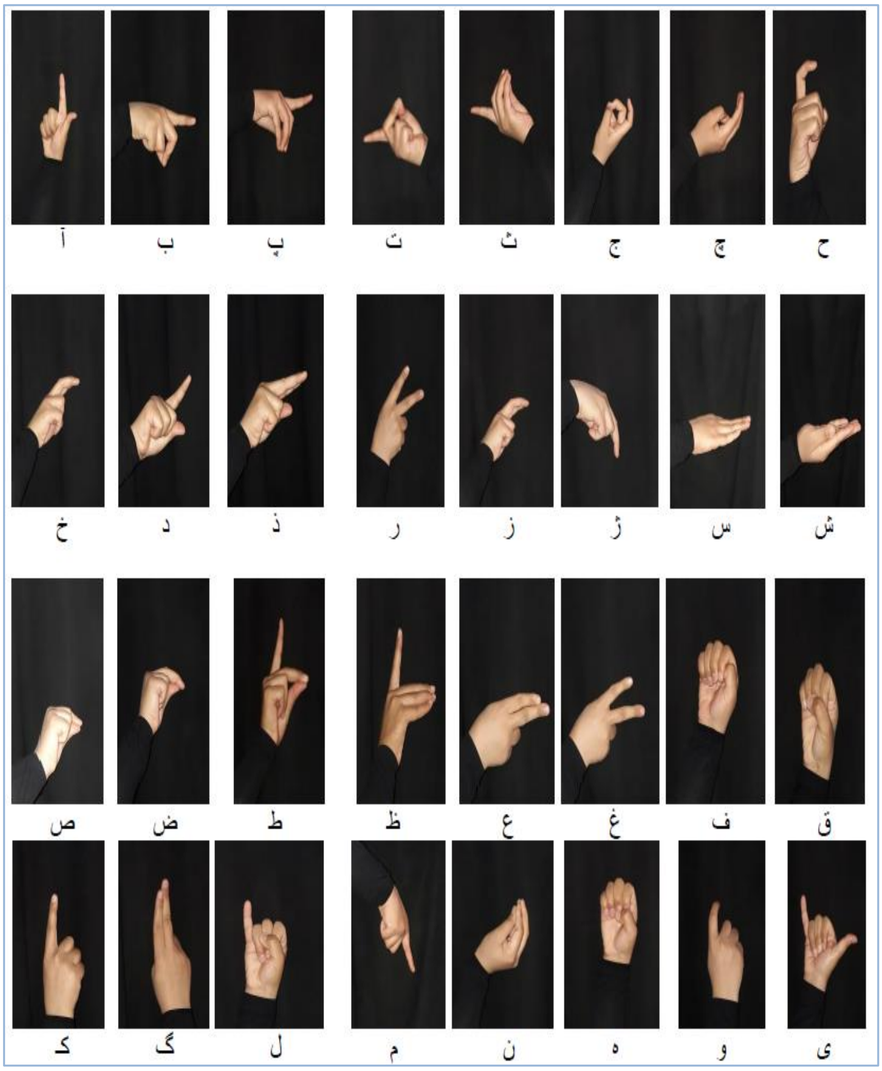
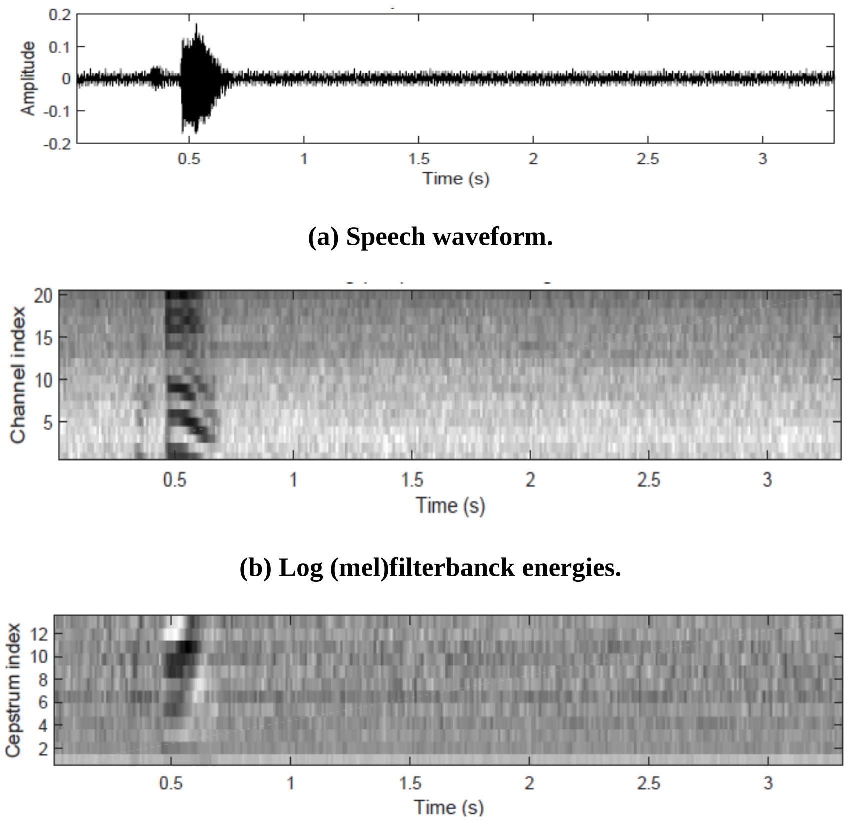


Fig. 2. Persian alphabet set [15].

stage, while the rest letters as shown in Appendix A, where the results of the classification are giving 100% accuracy in both LDA and QDA as demonstrated in Table 1 and Fig. 4 by using Eq.(10) [16].

$$\text{Accuracy} = \left(\frac{TP+TN}{TP+FP+FN+TN}\right)100\% \tag{10}$$

where TP Contains True Positives that properly classify positive states; FP is False Positives that improperly classify negative situations, TN is True Negative that properly classify negative states and FN is False Negative that is improperly classify positive states. The outcome from the testing stage gave (96.8%) by utilizing the Euclidian distance as demonstrated in Table 2.



(a) Speech waveform.

(b) Log (mel)filterbanck energies.

(c) Mel frequency cepstrum.

Fig. 3. The features extracted from voice letter(أ).

Table 1. The difference between two algorithms rate.

	No sound	LDA	QDA
Letter	128	100%	100%

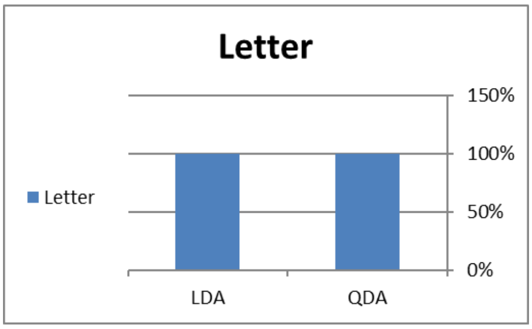


Fig. 4. The difference between two algorithms rate.

Table 2. The accuracy rate using Euclidian distance.

No. of tested sounds	**No. of dataset sounds**	**Euclidean distance**
32	128	96.875%

6. Conclusion

In this paper, the proposed system results show that highlight feature extraction from MFCC is superiorly better than LPCC, and LPC. The result outcome also shows that QDA and LDA preferred precision over different classifiers ways. Converting voice into gesture may make the communication between the blind people who can just speak (where they can speak out using their voice) to communicate with the dumb people they can understand the blind by gestures (images). This is the key to our proposed system.

Nomenclatures

m	$0 \le m \le M - 1$
N	The number of samples in each frame
u	0,1,2, . . ., *N*-1
v	0,1,2, . . ., *M*-1
W(m)	The window function
X(x,y)	Input signal *x*=0,1,2,. . ., *N*-1 and y = 0,1,2,. . ., *M*-1

Greek Symbols

α	A pre-emphasis filter constant, and the value usually set as $0.9 < \alpha < 1.0$. q_t is a frequency energy **.**

Abbreviations

DCT	Discrete Cosine Transform
FFT	Fast Fourier Transform
LDA	Linear Discriminate Analysis
LPCC	Linear Predictive Cepstral Coefficients
LPC	Linear Predictive Coding
MFCC	Mel Frequency Cepstral Coefficients
QDA	Quadratic Discriminate Analysis

References

1. Sears, J.T.; and Goldberg, D.A. (1998). *Voice-output reading system with gesture-based navigation*. *Ascent Technology*, Inc., Boulder, European Patent Office (EP1050010A1).
2. Swathy, S.M.; and Mahesh, K.R. (2017). Review on feature extraction and classification techniques in speaker recognition. *International Journal of Engineering Research and General Science*, 5(2), 78-83.
3. Alim, S.A.; and Rashid, N.K.A. (2018). Some commonly used speech feature extraction algorithms. *From Natural to Artificial Intelligence-Algorithms and Applications*. United Kingdom: IntechOpen Limited.

4. Endah, S.N.; Adhy, S.; and Sutikno. (2017). Comparison of feature extraction MFCC and LPC in automatic speech recognition for Indonesian. *TELKOMNIKA*, 15(1), 292-298.

5. Morais, C.L.M.; and Lima, K.M.G. (2018). Principal component analysis with linear and quadratic discriminant analysis for identification of cancer samples based on mass spectrometry. *Journal of the Brazilian Chemical Society*, 29(3), 472-481.

6. Assefa,B.G. (2012). *Non-uniform sampling based feature extraction for automatic speech recognition.* M.Sc. thesis, Addis Ababa University, India.

7. Muda, L.; Begam, M.; and Elamvazuthi, I. (2010). Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. *Journal of Computing*, 2(3), 138-143.

8. Wahyuni, E.S. (2017). Arabic speech recognition using MFCC feature extraction and ANN classification. *Proceeding of the 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering*. Yogyakarta, Indonesia, 22-25.

9. Vijayan, A.; Mathai, B.M.; Valsalan, K.; Johnson, R.R.; Mathew, L.R.; and Gopakumar, K. (2017). Throat microphone speech recognition using MFCC. *Proceeding of the International Conference on Networks and Advances in Computational Technologies*. Thiruvananthapuram, India, 392-395.

10. Joshi, D.D.; and Zalte, M.B. (2013). Recognition of emotion from Marathi speech using MFCC and DWT algorithms. *International Journal of Advanced Computer Engineering and Communication Technology*, 2(2), 59-63.

11. Meseguer, N.A. (2009). *Speech analysis for automatic speech recognition*. M.Sc. thesis, Norwegian University of Science and Technology, Norway.

12. Tirumala, S.S.; Shahamiri, S.R.; Garhwal, A.S.; and Wang, R. (2017). Speaker identification features extraction methods: A systematic review. *Expert Systems with Applications*, 90, 250-271.

13. Fisher, R.A.; D, Sc.; and S, F.R. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2), 179-188.

14. Hastie, T.; Tibshirani, R.; and Friedman, J. (2008). *The elements of statistical learning (2nd ed.)*. Switzerland: Springer, 106-119.

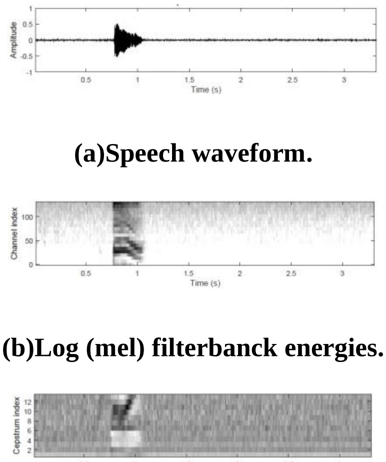
15. Karami, A.; Zanj, B.; and Sarkaleh, A.K. (2011). Persian sign language (PSL) recognition using wavelet transform and neural networks. *Expert Systems with Applications*, 38(3), 2661-2667.

16. Sheha, M.A.; Mabrouk, M.S.; and Sharawy, A. (2012). Automatic detection of melanoma skin cancer using texture analysis. *International Journal of Computer Applications*, 42(20), 22-26.

Appendix A

Representation and Figures of Design Charts

Appendix A lists the figures of results for letters (ب,پ ,ت,ث)

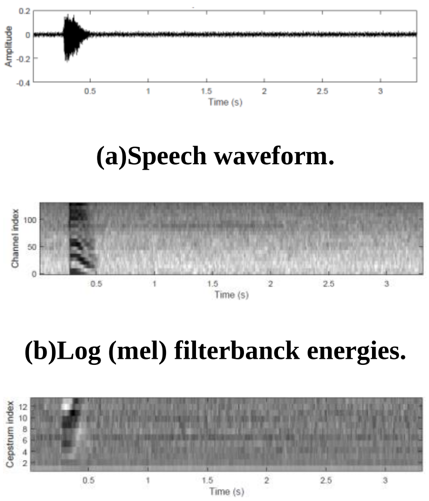


(a)Speech waveform.

(b)Log (mel) filterbanck energies.

(c)Mel frequency cepstrum.

Fig. A-1. the features extracted from voice letter(ب).

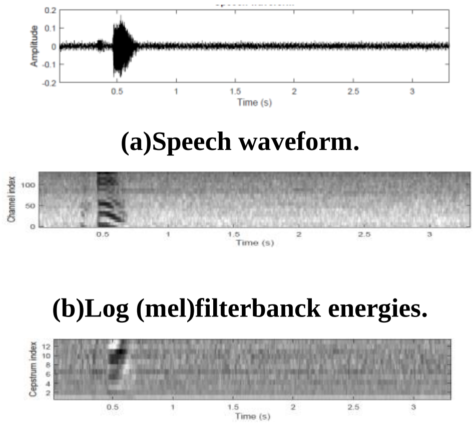


(a)Speech waveform.

(b)Log (mel) filterbanck energies.

(c)Mel frequency cepstrum.

Fig. A-2. the features extracted from voice letter(پ).

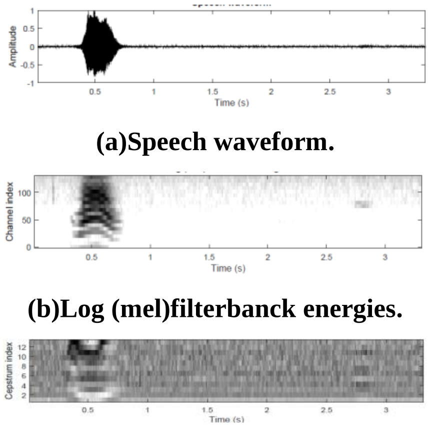


(a)Speech waveform.

(b)Log (mel)filterbanck energies.

(c)Mel frequency cepstrum.

Fig. A-3. the features extracted from voice letter(ت).



(a)Speech waveform.

(b)Log (mel)filterbanck energies.

(c)Mel frequency cepstrum.

Fig. A-4. the features extracted from voice letter(ث).

Appendix B

Computer Programme

In this Appendix will shows the proposed system diagram programme is shown in Fig. B-1.

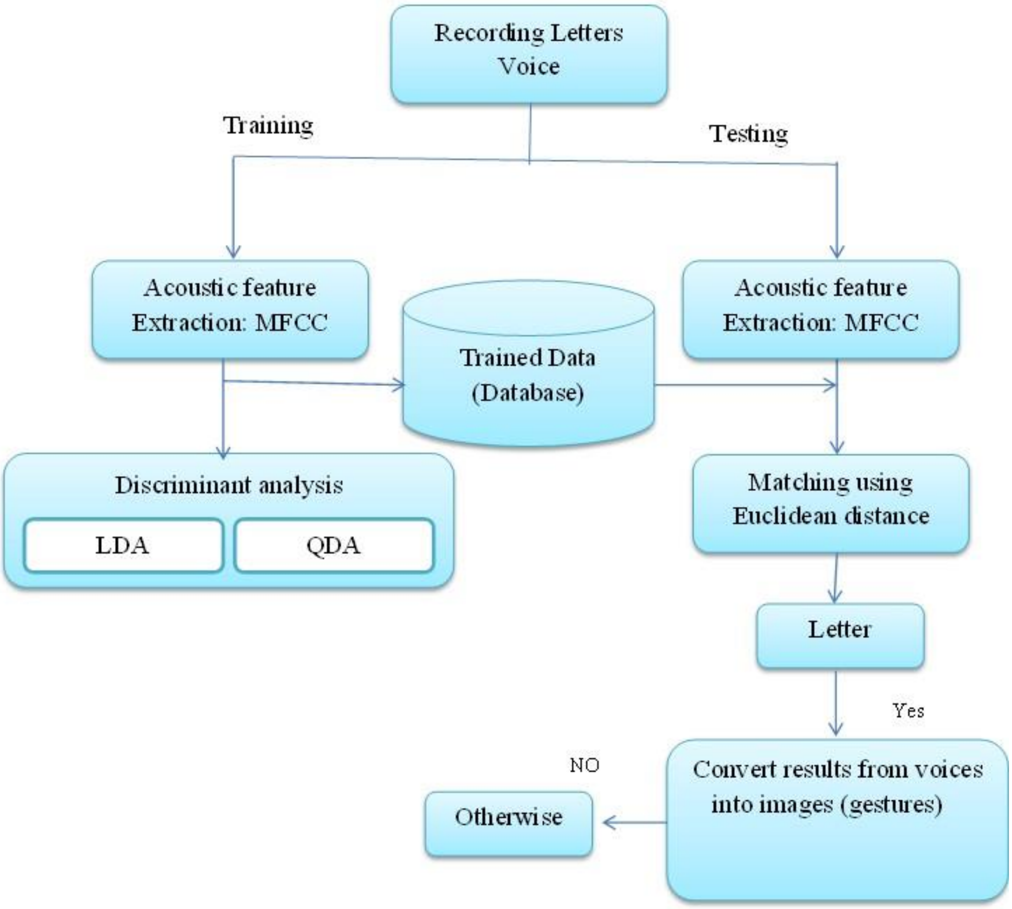

Recording Letters Voice
Training
Testing
Acoustic feature Extraction: MFCC
Trained Data (Database)
Acoustic feature Extraction: MFCC
Discriminant analysis
LDA
QDA
Matching using Euclidean distance
Letter
Yes
NO
Otherwise
Convert results from voices into images (gestures)

Fig. B-1. Proposed system.