Biomedical Research 2013; 24 (2): 185-192 ISSN 0970-938X

A robust speech disorders correction system for Arabic language using visual speech recognition.

Ahmed Farag[1], Mohamed El Adawy[2], Ahmed Ismail[3]

[1]Biomedical Engineering Department, HelwanUniversity, Egypt.
[2]Biomedical Engineering Department, HelwanUniversity, Egypt.
[3]Biomedical Engineering Department, HTI,Egypt.

Keywords: Speech Processing, Visual Speech, Arabic Speech Recognition, Speech Disorders Classification and Lips Detection.

Accepted January 02 2013

Abstract

In this Paper, we propose an automatic speech disorders recognition technique based on both speech and visual components analysis. First, we performed the pre-processing steps required for speech recognition then we chose the Mel-frequency cepstral coefficients (MFCC's) as features representing the speech signal.On the other hand, we studied the visual components based on lipsmovements analysis. We propose a new technique that integrates both the audio signal and the video signal analysis techniques for increasing the efficiency of the automated speech disorders recognition systems. The main idea is to detect the motion features from a series of lipsimages. A new technique for lips movement detection is proposed. Finally we use the multilayer neural network as a classifier for both speech and visual features.We propose a new technique for speech disorders correction systems, especially for Arabic language. Practical experiments showed that our system is useful when dealing with Arabic language speech disorders.

1- Introduction

In our daily communications, humans identify speakers based on a variety of attributes of the person which include acoustic cues, visual appearance cues and behavioural characteristics (such as characteristic gestures, and lip movements). In speech disorders classification the specialist depend on both the speech information and how to treat these artifacts through the visual information; much of the speech information is retrieved from the visual clues. In the past, machine implementations of person identification have focused on single techniques relating to audio cues alone (speaker recognition), visual cues alone (face identification, iris identification) or other biometrics. Automatic speech recognition (ASR), referred to as automatic lip-reading, introduces new and challenging tasks compared to traditional audio-only ASR. ASR uses the images sequence segmented from the video of the speaker's lips, which is the technique of decoding speech content from visual clues such as the movement of the lip, tongue and facial muscles. Automatic speech recognition (ASR) has recently attracted significant interest of many researchers [1, 2, 3].

Much of this interest is motivated by the fact that the visual modality contains some complementary information to the audio modality [4], as well as by the way that human fuse audio-visual stimulus to recognize speech [5, 6]. Not surprisingly, ASR has been shown to improve traditional audio-only ASR performance over a wide range of conditions Human lip reading Speech generally is multimodal in nature [1]. The human speech perception system fuses both acoustic and visual cues to decode speech produced by a talker. So, lip reading started when human started to know the language. In [7] the researchers found that lip information can lead to good improvement of human's perception of speech, this improvement goes better in a noisy environment.

The Paper is organized as follow: section two explains the proposed general block diagram of the system. This section covers the pre-processing of the speech segment to be prepared to the feature extraction step.These processes includes removing silence, pre-emphasis filter and at last windowing and framing step. It alsoincludesthe audio features extraction procedure.The more accurate this step is, the more accuracy at the classification step we can get. So we used the Mel-frequency cepstral coefficients MFCCs as a feature matrix for the speech segment. Section 3 covers the visual component analysis. It includes face detection, lips segmentation, and the geometry features extracted from the lips. Finally, section four covers the classification process by using multi-layer feed forward neural network with two output nodes (true, false) to indicate if the pronunciation is correct or incorrect, while section five discusses the results of classification for both normal and abnormal speech segments for the character "Raa" ("راء").Finally Section six contains a discussion about what have been achieved through this research and future work.

2- Proposed System

The System Input depend on two main signals: the first one is the interface with the microphone to get the speech segment into the computer as a replacement for the specialist ears, the second input is the webcam interface which gets the video motion of the lips as the mirror used by the specialist to get the correct movement from the patient.The PCcontains all the required processes by which the speech specialist brain's performs to make a decision whether the speech segment is correct or incorrect. Our experimental results showed that the proposed system is a unique system which depends on the speech segment supported by the visual movements that caused this speech.So after the output is released if one of the patterns was incorrect,either the audio or the visual pattern, the system will order the patient to repeat the speech again.This is to make sure that the patient is pronouncing the letter with the right pronunciation methodology and with the right lips movement. The proposed system is shown in Fig.1.

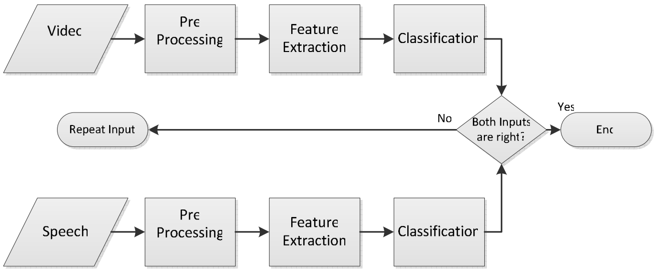


Figure 1. *The block diagram of the proposed system.*

Speech Pre-Processing

Before the feature extraction process, there are many required processes that must be done on the speech segment to be able to extract the MFCC's. We added a process at the first step to minimize the percentage of error during features extraction phase and consequently the classification phase. This step is called " removing silence" then the regular pre-processing steps for speech recognition systems were applied which are pre-emphasis filter, framing and windowing. All of pre-processing procedures are explained in Fig.2.

Removing Silence

This step was very important to reduce the percentage of error in feature extraction and increase the accuracy of classification and reduce the process time from about five seconds to just 1.4 seconds on the average.

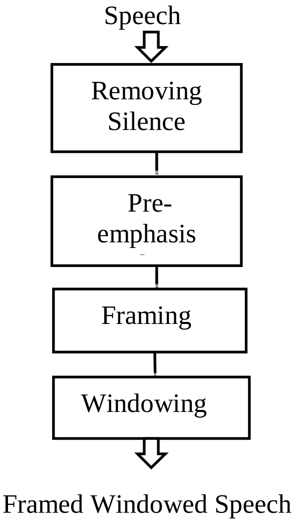


Figure 2. *The block diagram for pre-processing phase.*

To remove silence we depend on the signal energy and spectral centroid features of the audio signal.

The signal energy is defined as follows: Let x_i (n); n = 1…, N be the audio samples of the ith frame, of length N. Then, for each frame i the energy is calculated according to the equation:

$$E(i) = \frac{1}{N}\sum_{n=1}^{N}\left|x_i(n)\right|^2 \qquad (1)$$

This simple feature can be used for detecting silent periods in audio signals, and for discriminating between audio classes.

The spectral centroid is defined as follows: The spectral centroid, C_i, of the ith frame is defined as the centre of “gravity” of its spectrum, according to the following equation:

$$C_i = \frac{\sum_{k=1}^{N}(k+1)X_i(k)}{\sum_{k=1}^{N}X_i(k)} \; X_i(k) \qquad (2)$$

Where k=1,….,N , is the Discrete Fourier Transform (DFT) coefficients of the ith short-term frame, N is the frame length. This feature is a measure of the spectral position, with high values corresponding to “brighter” sounds [8].

Pre-Emphasis

Before the digital speech signal can be used for feature extraction, a process called pre-emphasisis applied. High frequency formants have lower amplitudes than low frequency formants. Pre-emphasis aims at reducing the high spectral dynamic range. Pre-emphasis is accomplished by passing the signal through an FIR filter [3] whose transfer function is given by:

$$H(z) = 1 - az^{-1} \quad where \ 0.9 \le a \le 1 \qquad (3)$$

A typical value for the pre-emphasis parameter ‘a’ is usually 0.95 [9].

Frame Blocking

The idea of segmentation of the speech wave into frames, or what is known as frame blocking, comes from the fact that the vocal tract moves mechanically slowly, and as a result, speech can be assumed to be a random process with slowly varying properties [10]. Hence, the speech can be divided into frames, over which the speech signal is assumed to be stationary with constant statistical properties. Another property must be guaranteed to ensure the continuity of the speech signal; this is generally done by overlapping the different frames. Note that typical values for the frame period are 45 ms with a 15 ms separation. This corresponds to a 16.7 Hz frame rate.

Feature Extraction

The extraction of the best parametric representation of acoustic signals is an important task to produce a better recognition performance. The efficiency of this phase is important for the next phase since it affects its behaviour. Mel-frequency cepstral coefficients MFCCs are based on human hearing perceptions which cannot perceive frequencies over 1KHz. In other words, MFCC is based on known variation of the human ear’s critical bandwidth with frequency. MFCC has two types of filter which are spaced linearly at low frequency below 1000 Hz and logarithmic spacing above 1000Hz. A subjective pitch is present on Mel Frequency Scale to capture important characteristic of phonetic in speech. The overall process of the MFCC is illustrated in Fig. 3.

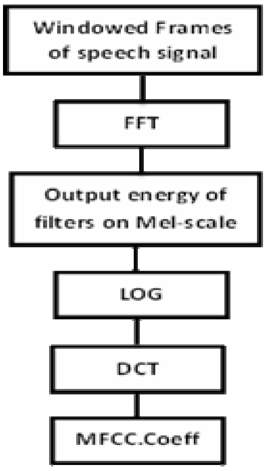

Windowed Frames of speech signal
FFT
Output energy of filters on Mel-scale
LOG
DCT
MFCC.Coeff

Figure 3. *The steps for calculating the MFCC.*

Face Detection Algorithm

The procedures of our component based algorithm for face detection is shown in Fig. 4.First, the RGB colour-space is transformed to the L,a,b colour space model. The skin tone pixels are extracted using the a and b components with ignorance of the luminanceeffect .The detected skin tone pixels are iteratively segmented into connected components by using colourvariance. After applying acombination of morphological operations, face candidate pixels are grouped. This helps us to extract the face from the whole picture with neglectingof the background. A mask is applied after this to crop the face as a ROI region from the image.This helpus to decrease the percentage of error in the following lips detection stages.

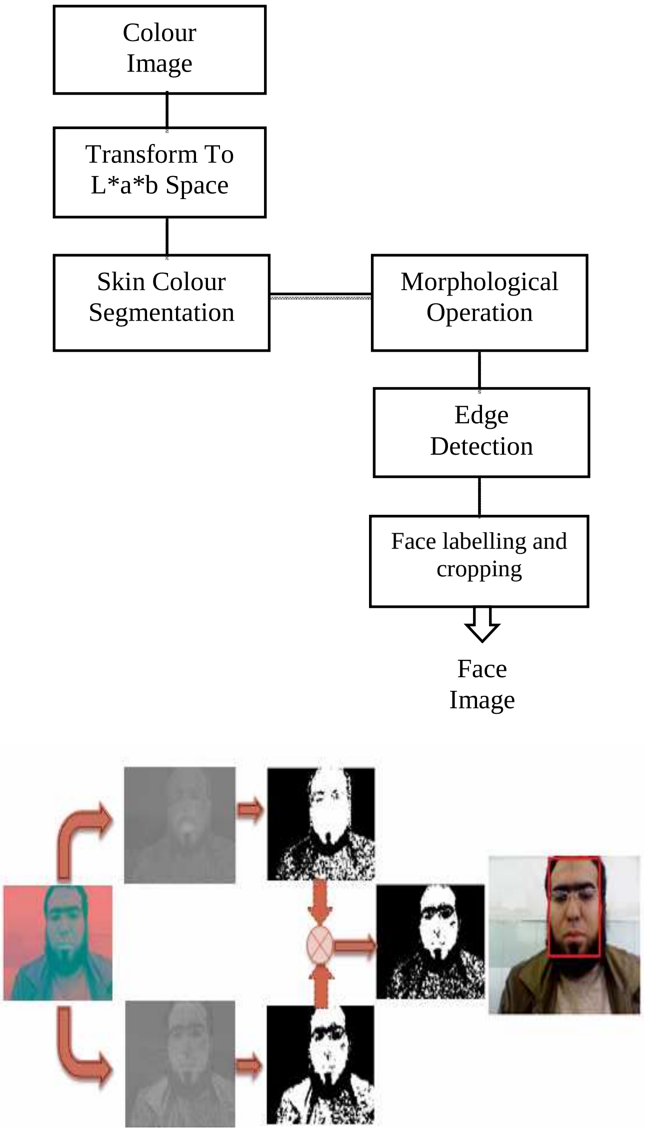


Figure 4. *Face detection technique using Lab colour space.*

Lips Segmentation

First the image is transformed from the RGB color space to YIQ space. Then, we extract the Q colourcomponent from the face image.

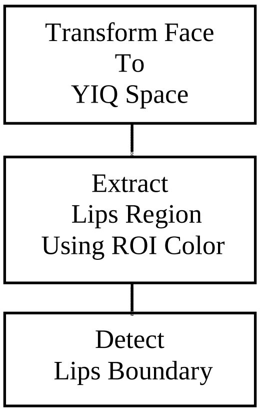
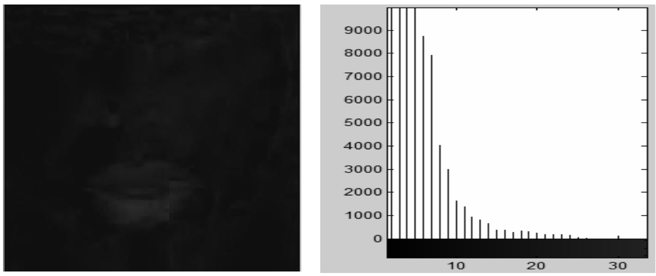


Figure 5. *Lips Segmentation.*

Then we calculate the optimum threshold for transforming the Q component image into binary form using the histogram technique as shown in Fig. 6.



(a) (b)

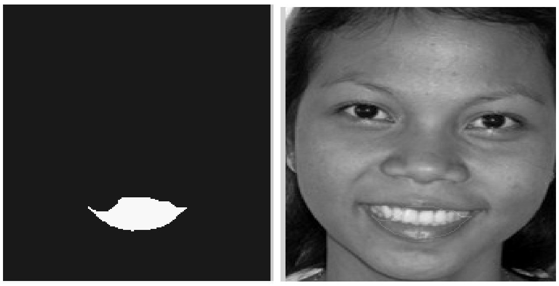
Figure 6.
(a) The grayscale level appearance for lips region
(b) Histogram distribution.

Lips Contour Extraction

The boundary of a set A, denoted as β(A), can be obtained by first eroding A by B and then performing the set difference between A and its erosion as follows:

$$\beta(A) = A - (A \Theta B) \qquad (5)$$

Where B is a suitable structuring element.



(a) (b)

Figure 7. *(a) Closed boundary of lips;*
(b) Boundary extraction applied to a real face.

We use the Implemented function in Matlab which is bwperim(); the function returns a binary image containing only the perimeter pixels of objects in the input image.

A pixel is part of the perimeter if it is nonzero and it is connected to at least one zero-valued pixel.

Visual Features

Visual features extraction depends mainly on the binary morphological operation. We consider morphological algorithms for extracting boundaries, connected components, the convex hull, and the skeleton of a region. Fig. 8 illustrates the proposed block diagram to have visual features that represent lips motion.

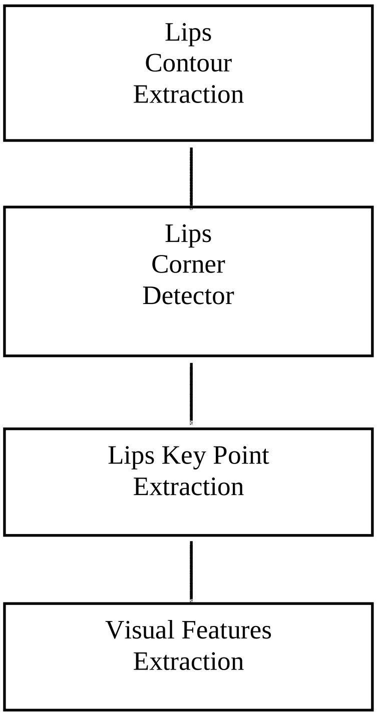


Figure 8. *Visual features extraction steps*

Lips Corner Detection

There are many methods for corner detection like Minimum Eigenvalue Method, Local Intensity Comparison and Harris Corner Detection Method .The Harris method advantage is the trade-off between accuracy and speed. The Harris method depend on the change of intensity for the shift [u,v], the following equation represent the Harris detector idea:

$$E(u,v) = \sum_{x,y} w(x,y)[I(x+u,y+v) - I(x,y)]^2 \qquad (6)$$

Where w(x, y): represent the window function.
I(x+u, y+v): represent the shifted intensity.
I(x, y): represent the intensity.

Harris detector measures the corner response R as shown in the following equation which depends only on eigenvalues of M and R which is large for a corner, R is negative with large magnitude for an edge, |R| is small for a flat region.

$$R = \det M - k(\operatorname{trace} M)^2 \qquad (7)$$

The Harris corner detection extracts about 9 points from the lips boundary.

Lips Key Point Extraction

We select four points to detect the change of these points during the speech of the same character with its different vowels. These four points are based on the maximum points in the lips corner at x and y direction. Then the minimum points in the lips corner at x and y direction.

Visual Features

There are primarily two categories of visual feature representation in the context of speech recognition. The first is model-based or geometric-based. Examples of such features are width and height of the lips (and their temporal derivatives) that can be estimated from the images. The second category is pixel-based or appearance-based; that is, the features are based on intensity values of the raw pixels. The first category is more intuitive, but there is typically a substantial loss of information because of the data reduction involved [11, 12].

We propose new features for general visual speech recognition. The extracted features are: The Maximum Distance according to X-direction, the Maximum Distance according to Y-direction and the area of mouth contour.

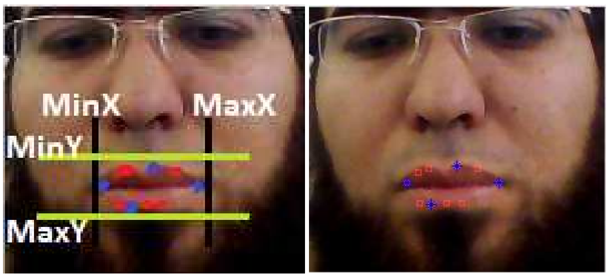
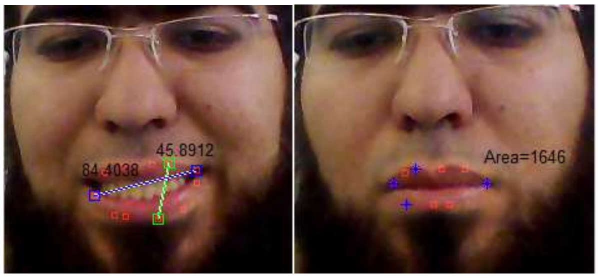


***Figure 9**. Key points for real lips boundary.*

We depend on distance measuring on the Euclidean distance theory. The Euclidean distance is the straight-line distance between two pixels as shown in Fig. 9 and derived by the equation (8):

$$dist((x,y),(a,b)) = \sqrt{(x-a)^2 + (y-b)^2} \qquad (8)$$



(a) (b)

***Figure 10.**(a) 2 maximum and minimum values for x, y directions for a real face, (b) The area of the lips boundary.*

Classifier

We used the neural network as a classifier .The neural network (NN) used in the model was a multilayer perceptron (MLP) with two layers of neurons. The number of neurons in the hidden layer depends on the size of the input vector [13]. The output layer has two neurons. The first neuron predicts if the input is a truly pronounced word or sentence. The second neuron predicts if the input is a wrongly pronounced word or sentence. The NN is trained to predict one true word or sentence at a time and whichever of these neurons gives the higher score wins. If an MLP network has n input nodes, one hidden-layer of m neurons, and two output neurons, the output of the network is given by:

$$y_i = f_i\left(\sum_{k=1}^{m} w_{ki} f_k\left(\sum_{j=1}^{n} w_{kj} x_j\right)\right) \qquad (9)$$

where f_k , k =1, 2,…m , and f_i, i= 1,2denote the activation functions of the hidden-layer neurons and the output neurons, respectively; w_{ki} and w_{kj}, j= 1,2,…,n denote the weights connected to the output neurons and to the hidden-layer neurons, respectively, and X_j denotes the input. The output activation function was selected to be unipolar sigmoidal:

$$f(u) = \frac{a}{1+e^{-\beta u}} \qquad (10)$$

And the hidden-layer activation functions took the form of hyperbolic tangent sigmoidal for all k:

$$f_k(u) = a_k \frac{e^{\beta_2 u} - e^{-\beta_2 u}}{e^{\beta_2 u} + e^{-\beta_2 u}} \qquad (11)$$

Neurons are updated iteratively. The weights and biases of the neural network were initialized as [-1.0, 1.0]. Desired outputs were set to either 0.9 or 0.1 to represent the true or false site at the output, correspondingly.

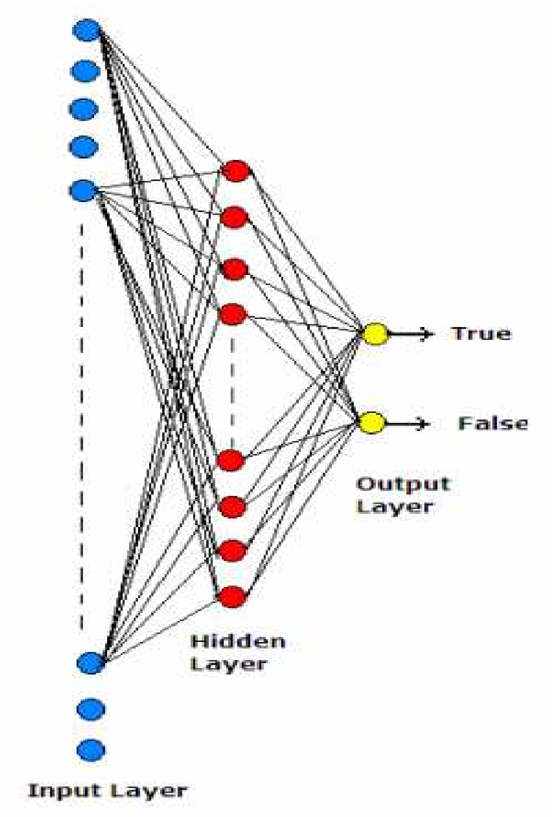


Figure 11.represent the architecture for multi-layer feed forward neural network.

Results

Wehave extracted two groups of features, one for each signal type. Each group of features represents a pattern to our classifier. The first pattern represents the speech segment and the second one represents the visual signal. The speech pattern represents the 8 MFCC's coefficients of the speech character as shown in figure 12.We utilized these coefficients in a single pattern that represents the MFCC's energy. On the other hand, the visual signal is represented by three features pattern. These features are x-distance, y-distance, and the area as shown in Fig.14.

Fig.12.a illustrates the MFCC coefficients for audio signal of normal speech for character “RA”, while Fig. 14.a illustrates the visual features for the video signal of normal speech for character “RA”.

Fig.12.b illustrates the MFCC coefficients for audio signal of abnormal speech for character “RA”, while Fig. 14.b illustrates the visual features for the video signal of abnormal speech for character “RA”.

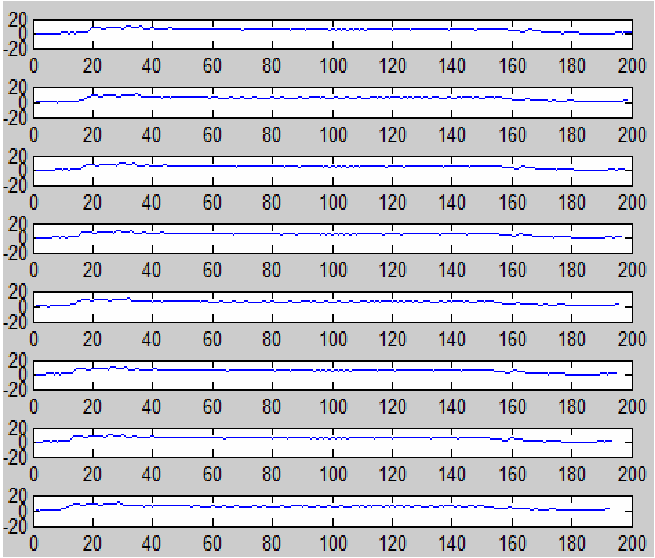


***Figure 12.a.** The 8 MFCC for normal speech for character " را ".*

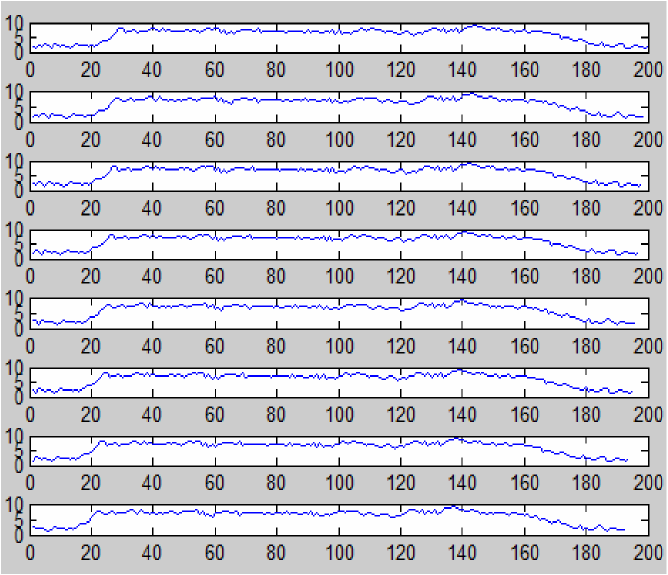


***Figure 12.b.** The 8 MFCC for abnormal speech for character " را "*

In Arabic language we usually have three different articulationsfor each character by adding the vowels to the same character for a lips movement during the pronunciation of the character (the vowels are: o, e, a). Fig. 13 displays the extraction of the lips boundary from stationary position and during the three movements for each articulation.

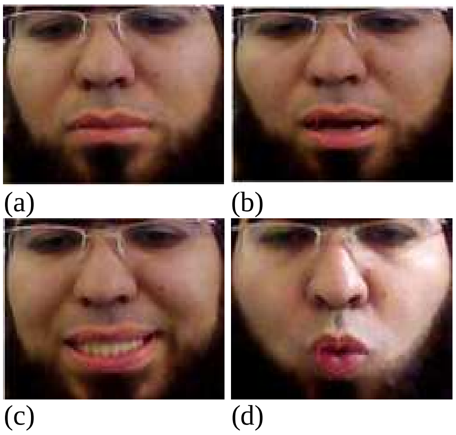

(a) (b) (c) (d)

Figure 13. *The lips boundary for pronunciation (a) stationary, (b) "Raa",(c) "Ree" ,(d) "Roo".*

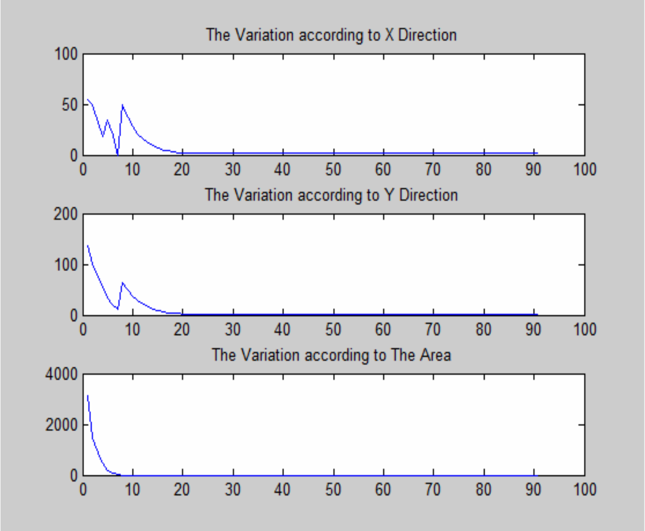


Figure 14a*. The 3 visual patterns for normal speech for character " ر ا ".*

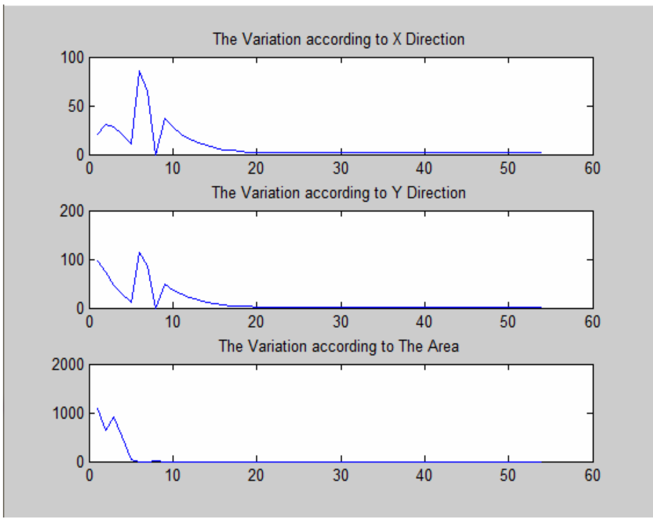


Figure 14b*. The 3 visual patterns for abnormal speech for character " ر ا ".*

The most important advantage of the proposed system is its dependence on both the speech signal and the visual movements detected from the video signal that caused this speech. So after the output is released if one of the patterns was fault this will order the patient to repeat the speech again. This guarantee that the patient has uttered the right pronunciation with right lips movement.
Table 1 shows the logical input output relationship of the proposed system.

Table 1. *Input/ output relationship*

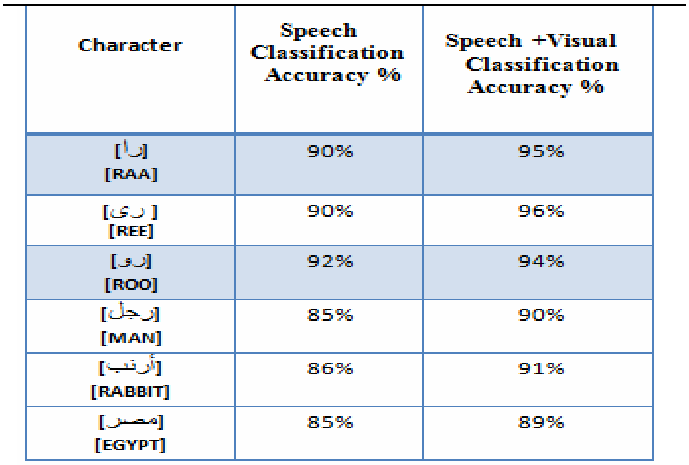
Inputs		**Output**
Speech pattern	**Visual pattern**	**Output**
False	False	False
False	True	False
True	False	False
True	True	True

Experimental results has shown that the accuracy of the proposed system has reached 95% in case of a single letter, while it has not been lower than 89% for a complete word as shown in Table 2. This accuracy is an acceptable accuracy by the experts in the field of speech disorders curing centres.

We have applied the proposed system in two different speech disorders curing centres. The first one was located in 10th of Ramadan city Cairo, Egypt, while the second one was located in Mansoura city.

Experts at both centres have accepted the results of the research and showed a great interest in using such an automated system to assist them in the curing process for speech disorders.

Table 2. *Classification Accuracy of the proposed system*



Character	Speech Classification Accuracy %	Speech +Visual Classification Accuracy %
[را] [RAA]	90%	95%
[رى] [REE]	90%	96%
[رو] [ROO]	92%	94%
[رجل] [MAN]	85%	90%
[أرنب] [RABBIT]	86%	91%
[مصر] [EGYPT]	85%	89%

Conclusion

In This Paper, an automated speech disorder correction system was presented for Arabic language using isolated word recognition system. The research was performed on letter " راء" with three vowels movement of this character and different position for the same character in the word (start of the word, middle of the word, and end of the word). The advantage of the system is that it depends on both speech and visual features to classify the speech signal which is very important for the speech disorder's treatment. Finally we usedthe feed forward multi-layer neural network to classify whether the pronunciation of this speech was correct or incorrect.
Future work can be done to enhance the classification accuracy and also to reduce the system complexity.

Acknowledgment

The authors greatly appreciate the support and cooperation of Speech disorders centres in 10th of Ramadan city and Mansoura city which allow collecting the database for a lot of children and the greet appreciation for the open source CVL database.

References

1. Stork DG, Hennecke ME, Speech reading by Humans and Machines, Berlin, Germany: Springer, 1996.
2. Teissier P, Robert-Ribes J, Schwartz JL, and A. Guerin-Dugue A. Comparing models for audiovisual fusion in a noisy-vowel recognition task," IEEE Transactions on Speech and Audio Processing, 7(6): 629-642, 1999.
3. Dupont S, Luettin J, "Audio-visual speech modeling for continuous speech recognition," IEEE Transactions on Multimedia 2000; 2(3): 141-151.
4. Massaro DW, Stork DG. "Speech recognition and sensory integration," American Scientist 1998; 86(3): 236-244.
5. H. McGurk and J.W. MacDonald, "Hearing lips and seeing voices," Nature, 264:746-748, 1976 .
6. A.Q. Summerfield, "Some preliminaries to a comprehensive account of audio-visual speech perception," In Dodd, B. and Campbell, R. (Eds.), Hearing by Eye: The Psychology of Lip-Reading. Hillside, NJ: Lawrence Erlbaum Associates, pp. 97-113. 1987 .
7. W. H. Sumby and I. Pollack, "Visual contributions to speech intelligibility in noise," Journal of the Acoustical Society of America, 26:212–215, 1954.
8. T. Giannakopoulos, “Study and application of acoustic information for the detection of harmful content, and fusion with visual information,” Ph.D. dissertation, Dpt of Informatics andTelecommunications, University of Athens, Greece, 2009.
9. Mohamad Adnan Al-Alaoui "Speech Recognition using Artificial Neural Networks and Hidden Markov Models" ,16-18 April 2008, Amman, Jordan
10. LotfiSalhi, TalbiMourad, and AdneneCherif "Voice Disorders Identification Using Multilayer Neural Network",,The International Arab Journal of Information Technology, Vol. 7, No. 2, April 2010.
11. Xuedong Huang, Alex Acero and Hsiao-Wuen Hon, Spoken Language Processing: A Guide to Theory, Algorithm and System Development, Carnegie Mellon University, pages: 230-231, April 2001.
12. G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior, “Recentadvances in the automatic recognition of audiovisual speech,” Proc.IEEE, vol. 91, no. 9, pp. 1306–1326, Sep. 2003.
13. A. Pinkus A. Approximation theory of the MLP model in neural networks", tech. rep., 1999.

Correspondence to:

Ahmed Farag
Department of Biomedical Engineering
HelwanUniversity
Egypt.