Enhancement Quality and Accuracy of Speech Recognition System Using Multimodal Audio-Visual Speech signal

Eslam E. El Maghraby*[1], Amr M. Gody *[2], M. Hesham Farouk **[3]

* *Electrical Engineering, Faculty of Engineering, Fayoum University Egypt*

** *Engineering Math. & Physics Dept., Faculty of Engineering, Cairo University Egypt*

[1]eem00@fayoum.edu.eg

[2]amg00@fayoum.edu.eg

[3]mhesham@eng.cu.edu.eg

Abstract— ***Most developments in speech-based automatic recognition have relied on acoustic speech as the sole input signal, disregarding its visual counterpart. However, recognition based on acoustic speech alone can be afflicted with deficiencies that prevent its use in many real-world applications, particularly under adverse conditions. The combination of auditory and visual modalities promises higher recognition accuracy and robustness than can be obtained with a single modality. Multimodal recognition is therefore acknowledged as a vital component of the next generation of spoken language systems. This paper aims to build a connected-words audio visual speech recognition system (AV-ASR) for English language that uses both acoustic and visual speech information to improve the recognition performance. Initially, Mel frequency cepstral coefficients (MFCCs) have been used to extract the audio features from the speech-files. For the visual counterpart, the Discrete Cosine Transform (DCT) Coefficients have been used to extract the visual feature from the speaker's mouth region and Principle Component Analysis (PCA) have been used for dimensionality reduction purpose. These features are then concatenated with traditional audio ones, and the resulting features are used for training hidden Markov models (HMMs) parameters using word level acoustic models. The system has been developed using hidden Markov model toolkit (HTK) that uses hidden Markov models (HMMs) for recognition. The potential of the suggested approach is demonstrated by a preliminary experiment on the GRID sentence database one of the largest databases available for audio-visual recognition system, which contains continuous English voice commands for a small vocabulary task. The experimental results show that the proposed Audio Video Speech Recognizer (AV-ASR) system exhibits higher recognition rate in comparison to an audio-only recognizer as well as it indicates robust performance. An increase of success rate by 4% for the grammar based word recognition system overall speakers is achieved for speaker independent test.***

Keywords- ***AV-ASR, HMM, HTK, MFCC, DCT, PCA, MATLAB, GRID.***

1 INTRODUCTION

Automatic speech recognition (ASR) is currently used as an assistive tool in many fields including human computer interfaces, telephony, and robotics and has been used as an alternative method for individuals with disabilities. In spite of their effectiveness, speech recognition technologies still need more work to be employed for people with speech communication disorder especially for people who find it difficult to type with a keyboard.

In human-human communication signals from multiple channels are at work. Human communicate not only through words but also by intonation, gaze, hand and body gestures and facial expressions. Human computer interaction can benefit from modeling several modalities in analogous ways. Multimodal systems represent and manipulate information from different human communication channels at multiple levels of abstraction. So, the need to other source of information that is related to speech can introduce a novel solution compared to audio only ASR. Visual features like the movement of the lips and facial features can work as an example of such source of information. Visual features are demonstrated in many recent audio-visual ASR systems for normal speakers [1, 2].

Hearing impaired and deaf persons make extensive use of visual speech cues and some few individuals perform lip-reading to such a degree that enables almost perfect speech perception [3]. It is well known that seeing the talker's face in addition to hearing his voice can improve speech intelligibility, particularly in noisy environments [4], [5]. The main advantage of the visual signal is its complementarity to the acoustic signal [6]. Phonemes that are most difficult to perceive in the presence of noise are easier to distinguish visually and vice versa. The visual signal contains that kind of information that is acoustically most sensitive to noise [3]. Studies have also shown that visual information leads to more accurate speech perception even in noise-free environments [7]. The strong influence of visual speech cues on human speech perception is demonstrated by the McGurk effect [8] in which, for example, a person hearing an audio recording of /baba/ and seeing the synchronized video of a person saying /dada/ often resulted in perceiving /gaga/.

Automatic speech recognition (ASR) has been an active research area for several decades, but in spite of the enormous efforts, the performance of current ASR systems is far from the performance achieved by humans: error rates are often one order of magnitude a part [9]. Most state-of-the-art ASR systems make use of the acoustic signal only and ignore visual speech cues. They are therefore susceptible to acoustic noise [10], and essentially all real-world applications are

subject to some kind of noise. Much research effort in ASR has therefore been directed toward systems for noisy speech environments and the robustness of speech recognition systems has been identified as one of the biggest challenges in future research [11].
The advantage of such an approach is straightforward; the weaknesses of one modality are offset by the strengths of another, resulting in higher accuracy levels. Indeed, audio-visual speech recognition (AV-ASR), in which acoustic features and visual information extracted from the speaker mouth region are jointly used, has been investigated in the literature and found to increase ASR accuracy, primarily in the presence of acoustic noise [12,13].

The above facts have motivated significant interest in automatic recognition of visual speech, formally known as automatic lip reading, or speech reading [5]. Work in this field aims at improving ASR by exploiting the visual modality of the speaker's mouth region in addition to the traditional audio modality, leading to audio-visual automatic speech recognition systems. Critical however to the performance of the resulting audio-visual ASR system is the choice of visual features that contain sufficient information about the uttered speech.

There are three key reasons why vision benefits human speech perception [14]: It helps speaker (audio source) localization, it contains speech segmental information that supplements the audio, and it provides complimentary information about the place of articulation. The latter is due to the partial or full visibility of articulators, such as the tongue, teeth, and lips. Place of articulation information can help disambiguate, for example, the unvoiced consonants /p/ (a bilabial) and /k/ (a velar), the voiced consonant pair /b/ and /d/ (a bilabial and alveolar, respectively), and the nasal /m/ (a bilabial) from the nasal alveolar /n/ [15]. All three pairs are highly confusable on basis of acoustics alone. In addition, jaw and lower face muscle movement is correlated to the produced acoustics [16–17], and its visibility has been demonstrated to enhance human speech perception [18].

Compared to audio-only speech recognition, AV-ASR introduces new and challenging tasks, that are highlighted in the block diagram of Figure1: First, in addition to the usual audio front end (feature extraction stage), visual features that are informative about speech must be extracted from video of the speaker's face. This requires robust face detection, as well as location estimation and tracking of the speaker's mouth or lips, followed by extraction of suitable visual features. In contrast to audio-only recognizers, there are now *two* streams of features available for recognition, one for each modality. The combination of the audio and visual streams should ensure that the resulting system performance is better than the best of the two single modality recognizers, and hopefully, significantly outperform it. Both issues, namely the *visual front end design* and *audio-visual fusion*, constitute difficult problems [19], and they have generated much research work by the scientific community.

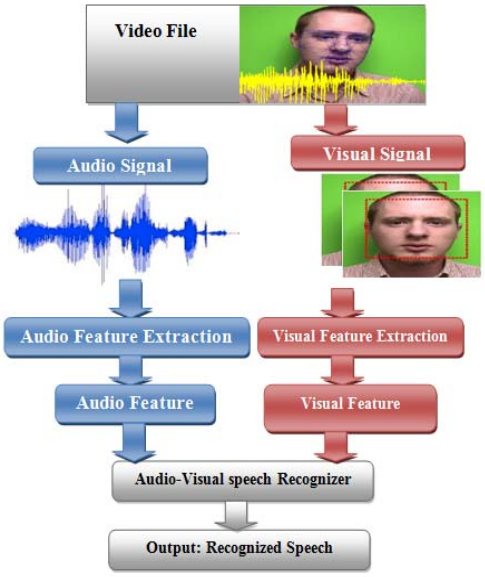


Figure 1: audio-visual speech recognition system

The accuracies obtained by the previous researches in Audio visual speech recognition system are reasonably high [19], but it is still needed to get further improvement. This paper describes a system that uses the visual features to enhance the recognition accuracy.
The proposed Audio Video Automatic Speech Recognizer (AV-ASR) system extracts solely appearance based features, and operates on full face video with no artificial face markings. As a result, both face detection and ROI extraction are required. All stages of the adopted visual front end algorithm are described below.

This paper will discuss the effect of adding visual features on the performance of speech recognition system for different visual features selection methods compared to audio only speech recognition systems and give result on English sentence corpus.

The rest of this paper is organized as follows. Section 2 discusses previous related works. Section 3 explains the block diagram of our proposed system. The experimental results are introduced in section 4. Section 5 contains a conclusion about what have been achieved through this research and future work.

2 PROPOSED SYSTEM

The proposed AV-ASR system architecture that is introduced in this paper is depicted in Figure 2. The input video that contains the speakers' spoken word is divided into audio file and its corresponding image files. There are two working threads, the audio front-end and images (visual) front-end. The audio-visual feature integration process is then performed. Finally, the Hidden Markov Model (HMM) classification is applied to classify the words to their respective classes.

A. Audio Front-End

In this subsection, preprocessing steps done on the audio files and feature extraction are described.

1) *Audio alignment with video stream*: GRID corpus is used. The audio is extracted from the composite video signal. This is accomplished by using the following command to extract mono channel audio signal from the composite signal (mpg) file, run this command line in windows command bacth file,

for f in *.mpg; do ffmpeg -i "$f" -ac 1 "${f%.mpg}.wav"; done

2) *Audio Pre-Processing:* Before extracting the ASR features, there are required pre-processes that must be applied on the speech streams.
   - ***Framing***: or segmentation, means dividing the speech signal into smaller pieces to alter it as stationary with constant statistical properties. It is common in speech to use frame length window not more than 25(ms). In other words, speech signal holds its properties for small period of time typically 25ms [19].
   - ***Frame Overlapping:*** Another process that is optionally used to ensure the continuity of the speech signal properties in the current frame along with the adjacent frames. The typical value for the frame overlap period is 10ms [19].
   - ***Frame Scaling:*** Since Speech is a non-stationary signal where properties change quite rapidly over time. For most phonemes the properties of the speech remain invariant for a short period of time short-term which estimates of parameters and this is done by effectively cross-multiplying the signal by a window function which is zero everywhere except for the region of interest. Hamming window is applied on the current frame.

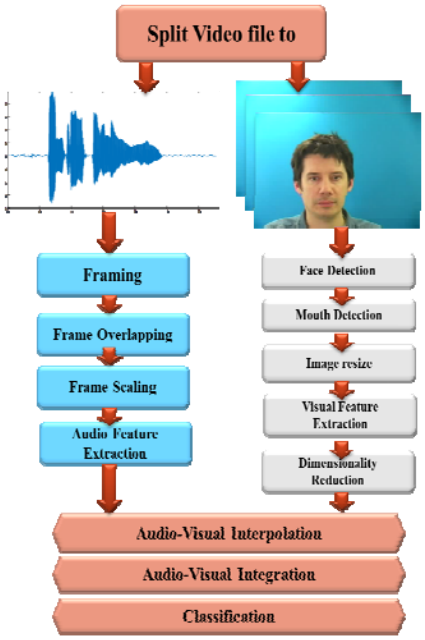


Figure 2: The block diagram of the proposed audio-visual speech recognition system.

3) *Recognition Feature Extraction from Audio signal:* Mel frequency cepstral coefficients (MFCC) is chosen in this research paper. MFCC is the most common audio features [20], MFCC is based on known variation of the human ear's critical bandwidth with frequency. The overall process of the MFCC is illustrated in figure 3 The software system Hidden Markov Model Toolkit (HTK) [2] is used for extracting 13 MFCC features together with their 1st and 2nd derivatives producing an acoustic feature vector of length 39 elements.

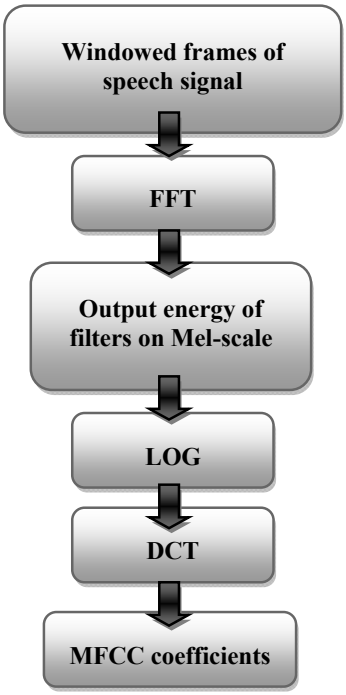


Figure 3: Steps of Calculating MFCC features.

B. *Visual Front-End*

The pre-processing on the images of the input video, visual feature extraction, and post processing are explained below. The mouth region within a rectangular window was detected as ROI. This was done by applying a classifier trained by the rapid and robust Viola-Jones object detection algorithm. These colored images are further transformed into gray-scale ones. By using appearance (pixel) based method, every pixel inside the detected ROI images was considered as a feature.

1) *Visual Pre-Processing*: We extract the visual features from mouth region; so, the Mouth region needs to be prepared first and this is done by some preprocessing steps which are briefly explained below:

- ***Face Detection*:** Before we begin tracking a face, we need to first detect it. Matlab [22] is used to detect speaker's face. It has tools for object detection like the *vision.CascadeObjectDetect*or to detect the location of a face in a video frame. The cascade object detector uses the Viola-Jones detection algorithm and a trained classification model for detection. A rectangle around the face region is returned.

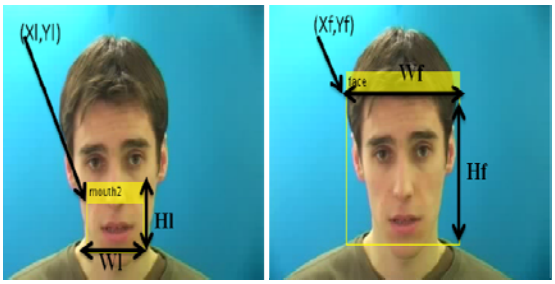
  The Visual Speech Recognition (VSR) system adopted the Viola-Jones detection module [23], which is much faster than any of its contemporaries via the use of an attention cascade using low feature number of detectors based on a natural extension of Haar wavelets. In this cascade, each detector fits objects to simple rectangular masks. In order to reduce the number of computations for such large number of cascades, Viola and Jones used the concept of integral image. They assumed that, for each pixel in the original image, there is exactly one pixel in the integral image, whose value is the sum of the original image values above and to the left. The integral image can be computed quickly, which drastically improves the computation costs of the rectangular feature models. The attention cascade classifiers are trained on a training set as the authors have explained in [24]. As the computation progresses down the cascade, the features can get smaller and smaller, but fewer locations are tested for faces until detection is performed. The same procedure is adopted for mouth detection, except that object is different and search about mouth will be only on the lower half of the input image.

TABLE 1

MOUTH LOCALIZATION ALGORITHM [25]

1. Grab the video frame for input.
2. Achieve the face detection and draw a box on the detected face then determined some of detection box (face) properties where:
   - the origin point
     - Xf: x-coordinate of the left border of face region
     - Yf: y-coordinate of the top border of face region
   - The width
     - Wf: the width of face region
   - The height
     - Hf: the height of face region
3. Detect the lip region is set as per the following calculations,
     - Xl=Xf+Wf/4
     - Yl=Yf+(2*Hf/3)
     - Wl=Wf/2
     - Hl=Hf/3

   Where:
     - Xl: x-coordinate of the left border of lip region
     - Yl: y-coordinate of the top border of lip region
     - Wl: the width of lip region
     - Hl: the height of lip region
4. Xl,Yl,Wl and Hl are the values constituting of the lip region in the lip detection.
5. Repeat step 2, 3 and 4 for all frames.



(b) (a)

Figure 4 The mouth localizing algorithm (a) the box around face (b) the box around mouth in proportional to the face box

- **Mouth Detection:** Mouth localization algorithm based on deciding the bounding box around the mouth in a geometric way. Proportional to the bounding box around the face, the algorithm decides first the left corner point of the mouths bounding box. Then, the required size of this box can be drawn easily to the extent that encloses any possible lip movements. But the method introduced in [30] doesn't give accurate mouth detection method, so we adjust the mouth region to be calculated from the following formula:

```
Xl=fbox1(1)+((fbox1(3))*0.3);
Yl =fbox1(2)+(0.73*(fbox1(4)));
W1= (fbox1(3))*0.4;
Hl = (fbox1(4))*0.27;
mnbox=[Xl Yl W1 Hl];
```

  Table 1 shows the proposed mouth localization algorithm which was explained in [25]. Figure 4 (a) and (b) show how the algorithm decides the box around the mouth in proportion to the box around the face. After getting the rectangle arround the mouth region we use imcrop Matlab function to crop the ROI around the mouth region.
- **Resize Mouth Region:** Mouth rectangle is resized to be in the form of 2^n where n is an integer. This operation is done in order not to make the calculation of DCT features be affected by the lip location in the input image. We choose the value of n to be 6 (64*64 pixels). Using Matlab *imresize* function is used to do this task.
- **Convert RGB to Grey:** The input mouth image is of RGB format. It is converted to grey format in range 0 (black) to 255 (white). We use *rgb2gray* function for implementing this process.

As simple output of each pre-processing step of visual front-end for an image from a speaker is shown in figure 5

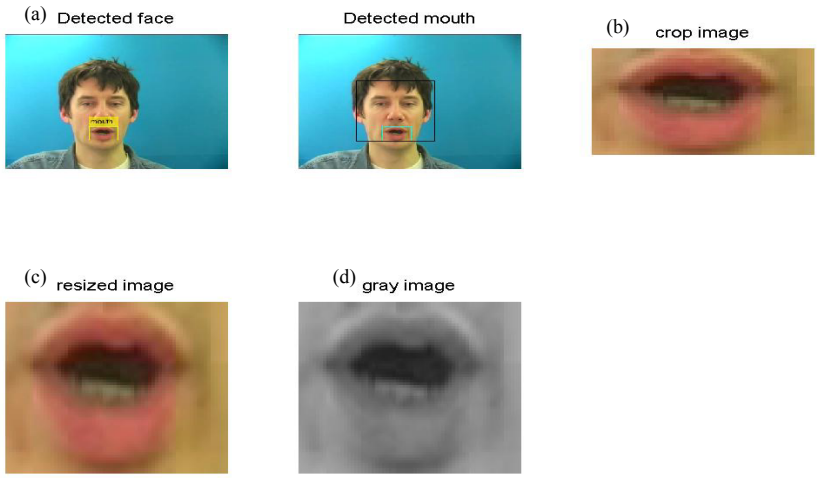


Figure 5: Visual pre-processing steps. a) Face and mouth detected regions, b) Mouth region only, c) Mouth region after resizing by 64x64, and d) Mouth region as gray scale.

C. Recognition Features Extraction for the visual signal

There are two main visual feature extraction categories that are appearance or pixel based and shape or model based. Examples of model based features are the width and the height of the speaker's lips. There is a loss of information

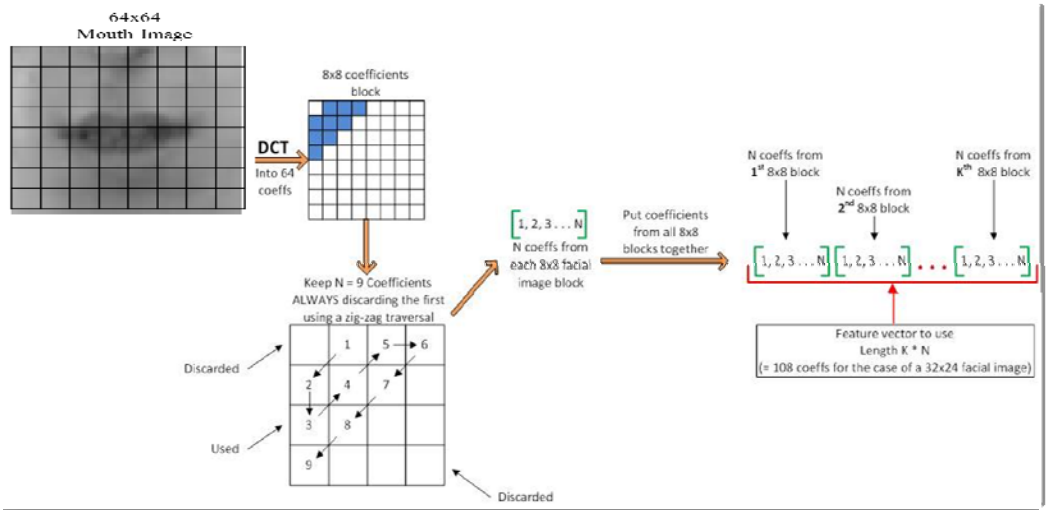
because it depends on some information about the lips not the whole region [26]. Appearance based assumes that all mouth region pixels are informative to speech recognition [27].

Various visual features have been proposed in the literature. In general, feature extraction methods can be categorized into three kinds: 1. "pixel based" where features are employed directly from the image, 2. "lip contour based", in which a prior template or model is used to describe the mouth area and 3. the combination of 1 and 2. Among these approaches, the one based on low level pixels is assumed to be the most efficient on [27]. As a typical method to extract pixel based features, image transforms such as Discrete Cosine Transform (DCT) [28], Principal Component Analysis (PCA) [29], Discrete Wavelet Transform (DWT) [28] and Linear Discriminant Analysis (LDA) [30] have been employed for lip-reading and have achieved high accuracy for visual-only recognition task. Among these, DCT has been shown to perform equally well or better than others.

Working at this pixel-based field faces a problem: How to reduce the high dimensional raw image data to low dimensional feature vectors without losing important information? Potamianos [29] retained the coefficients according to several sub lattices. Heckmann [32] compared 3 strategies to select the coefficients based on energy, variance and relative variance respectively and stated that the one based on energy performed best. Nefian [33] divided a 64 × 64 Region of Interest (ROI) into 64 blocks of size 8×8, and extracted the first 2x2 low frequency coefficients from each block. Projection using LDA to seek optimal classification performance in [33] can also be used for data dimensionality reduction, although this ability of LDA is limited to the number of classes. Motivated by the above studies, this paper focuses on the dimensionality reduction strategies for DCT based features for visual-only lip-reading task. In view of the excellent ability for information compression, PCA is applied to extract DCT coefficients. This combination is assumed to utilize the advantages of these two transforms. DCT is preferable to differentiate frequencies while PCA is beneficial to select the most 'important' components. Experimental results demonstrate that this new method does improve the speech reading performance when the final dimension is below a certain point, compared to the methods of selecting the coefficients according to specific criterion, such as 'low frequency'.

The visual feature extraction task concentrated in the current work. Inspired by the cascade strategy by [34],

- The first stage is the image transform by using block based DCT. This step forms a 320 dimension vector by using blocked 8x8 2D DCT and then extract 5 elements from the upper left corner from each block using zigzag method as shown in figure 6.
- The second step is the dimensionality reduction procedure, using PCA form a final vector V in the final stage. Then V is used as a feature vector of the visual part of the system.



- In PCA we take the first 9 eign vectors which have highest values, the dimension reduced from 320 to 9.

Figure 6: Extract feature vector from low frequency components from each block

D. Audio-Visual Features Integration

The features from different modalities (audio and visual features) have to be fused at some level. There are two d strategies to work with different types of features, early integration and late integration. In early integration (or called feature fusion), features from different sources are concatenated in one feature vector. The recognition pr applied on the combined feature vector. Late integration uses different or same classifiers for each feature type, a

the results of the classifiers are combined to get the final classification result. In this paper, early integration strategy is used by concatenating the acoustic and visual feature vectors on one vector. However, the audio and visual are with different frame rates, 44.1 KHz and 25 Hz for audio and video respectively, so linear interpolation is required first to up sample the video features rate to be with the same frame rate as audio features.

The Video features vectors are linearly distributed over the Audio features vectors to create the composite features vectors. The distribution is done by cloning the smaller set of Video features vectors in such that to build the same size array as such of the larger set Audio features vectors [18]. Then the composite features vectors are constructed by concatenating both arrays of features vectors (the cloned set video features vectors and the associated set Audio features vectors).

E. HMM Classification

Hidden Markov Model (HMM) is proven to be highly reliable classifier for speech recognition applications; most of the current successful systems for automatic speech recognition are based on Hidden Markov Models. The hidden Markov Model Toolkit by university of Cambridge (HTK) [21] is used for configuring, training and testing the HMM model are initialized using the Viterbi algorithm [35]. A total of 53 HMM models, one for each word, are trained in this paper. 44 phonemes are used to build each word's model. The proposed model uses 5-state left-to-right models with different number of Gaussian mixtures from 2 to 128 mixtures. Each state is multi Gaussian statistical model to express the observed symbols. In HTK, the conversion from single Gaussian HMMs to multiple mixture component HMMs is usually one of the final steps in building the model. The mechanism provided to do this is the HHED MU command which will increase the number of components in a mixture by a process called mixture splitting. This approach to building a multiple mixture component system is extremely flexible since it allows the number of mixture components to be repeatedly increased until the desired level of performance is achieved.

The MU command has the form: MU n itemList

where n gives the new number of mixture components required and itemList defines the actual mixture distributions to modify. This command works by repeatedly splitting the mixture with the largest mixture weight until the required number of components is obtained. The actual split is performed by copying the mixture, dividing the weights of both copies by 2, and finally perturbing the means by plus or minus 0.2 standard deviations e.g. MU 3 {*.state [2-4].mix}

It is usually increasing the number of mixtures then re-estimating, then incrementing by 1 or 2 again and re-estimating, and so on until the required numbers of components are obtained. This also allows recognition performance to be monitored to find the optimum mixture. Better start with a lesser number of mixtures and work way up. As one cannot go in the reverse direction, that is, there is no way to merge mixtures in HTK. So use single Gaussian models first then increment so as to reach a mixture of 8.

Performance analysis:

In order to analyze the system performance, HTK provides a tool HResult. It is used to compute the accuracy of the system. It compares the machine transcription of the test utterances with the corresponding reference transcription files. The performance of speech system is evaluated as:

$$\%Correct = \frac{N - D - S}{N} \times 100 = \frac{H}{N} \times 100 \qquad (1)$$

where N is the number of words in test set, D is the number of deletions, S is number of substitutions and H is the number of correct labels. %correct gives the percentage of word correctly recognized. The accuracy is computed as:

$$\%Accuracy = \frac{N - D - S - I}{N} \times 100 = \frac{H - I}{N} \times 100 \qquad (2)$$

where I is the number of insertions. The performance of speech recognition system can be evaluated by measuring the word error rate (WER) defined as:

$$\%Word\ Error\ Rate = \frac{S + I + D}{N} \times 100 = 100 - Accuracy \qquad (3)$$

3 EXPERIMENTAL RESULTS

This section presents the results of the experiments conducted to validate the effectiveness of the proposed design. Initially the performance of a baseline audio only recognition system is presented. Then, present the effect on the recognition accuracy of using visual features extracted from degraded video, and finally the result of audio visual system.

A. Data Description

To compare automatic audio visual speech recognition system performance based on the system discussed above; the GRID corpus [37] is used to perform these comparisons, which is a continuous audio-visual speech corpus for an English

small vocabulary task. It contains 1000 sentences spoken by each of 34 speakers (18 male, 16 female) ages ranged from18 to 49 years. The original audio and video data were recorded under clean acoustic conditions, and the video shows only a frontal view of each subject's face. The sentences in GRID are speech commands according to a very simple grammar. Each sentence in this database contained six words including a command, color, preposition, letter, digit, and adverb. The total of 51 words within the vocabulary consist of 4 command words, 4 words representing color, 4 prepositions, 26 letters, 10 digits and 4 adverbs. Example sentences produced by a speaker in this database were "bin blue at A1 again" or "place green by D2 now". The video was recorded as a sequence of images with a frame legnth of 40ms. In the audio channel, the raw speech signal was converted into a sequence of vector parameters with a fixed 25ms frame legnth. Table 2 and figure 7 introduce the grammar file for the GRID corpus.

TABLE 2

SENTENCE STRUCTURE FOR THE GRID CORPUS [37]

Command	Color	Preposition	Letter	Digit	Adverb
BIN LAY PLACE SET	BLUE GREEN RED WHITE	AT BY IN WITH	A-Z excluding W	1-9, zero	AGAIN NOW PLEASE SOON

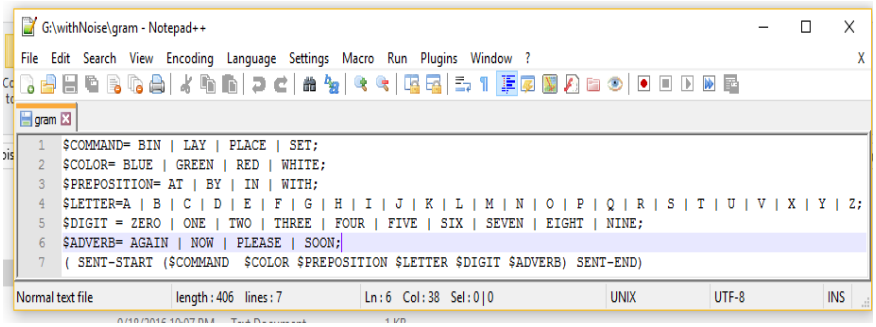


Figure 7: Grammar file for GRID corpus

- *The Experiment variables*

The experiment variables are listed below:

1- Number of the training and testing data used: We take 90% from the used database for training and 10% for testing
2- Audio Only Speech recognition system, Multimodal Speech recognition system.: We change the types of the feature used to check the improvement of adding the visual feature to the audio feature
3- Number of Gaussian Mixtures in HMM emitting states: We use 3 emitting HMM states and varying the numbers of the Gaussian mixtures from 2 to 128 to check if the increasing the numbers of the mixtures will increase the result or not.

First we take small amount of the database to make a small experiment so we take 100 files, 90 files for training and 10 for test.

- **Experiment** 1: speech recognition using acoustic features alone

In this experiment, we test audio only speech recognition system. HMM model with 3 emitting states and different Gaussian Mixtures in each state is used to model the recognition process. The average accuracy of using audio features only is summarized in figure 8 where the percentage correct against the experiment variables is represented, the qualifiers mono means monophone recognition, tri, means triphone recognition and Mix2 to Mix128 the number of Guassian mixtures increase from 2 to 128..

The parameters of each method used in figure 8 are explained in table 3 where A13 means Acoustic parameters with 13 MFCCs with 12 Mel cepstrum plus log energy and A39 mean the 13 elements with their delta (first order derivative) and acceleration (second order derivatives) coefficients.. Figure 8 proves that the audio only recognizer achieves high recognition rate for 39 feature vector size than using 13 elements for the feature vector, and the increase of mixtures number enhances the performance of recognition process for the two sizes of the feature vector. The optimal number of mixtures for A13 is 4 mixtures where it is 2 in A39. It gives us that increasing the feature vector size from 13 to 39 gives enhancement for the recognition rate by 6.6%.

TABLE 3

RESULT OF AUDIO ONLY SPEECH RECOGNITION SYSTEM

	Feature Type	Vector length	Best Result
A13	Audio only with MFCC_0	13	80% with Mix4
A39	Audio only with MFCC_0_D_A	39	86.67% with Mix2

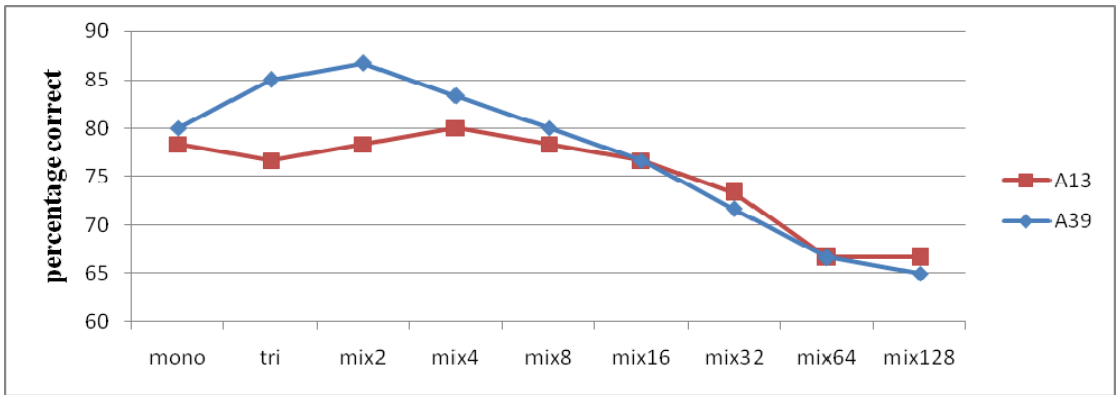


Figure 8: Percentage correct for audio only speech recognition with different mixture

- **Experiment** 2: speech recognition using Audio Visual features

The audio features and visual cues contain information related to speech production and combining these two signal streams can improve recognition accuracy; so we combined visual and acoustic features for dataset. In this experiment we check the change of the image resize effect from 64x64 and 128x128 the result is shown in table 3 which explains that resizing the image with 64x64 gives better results.

Change in the visual feature vector length from 2, 3, 5 and 6 the DCT matrix methods are applied and compared. The parameters of each method used in figure 9 are explained in table 3, where *AV45_128* means audio visual features with 45 feature vector size and 128x128 image size. The results indicate the best recognition obtained by using 64x64 image size with the DCT and PCA feature extraction. From table 4 we can see that best result is obtained when using audio visual feature with size of vector 42 and image size 64x64 with 91.67% recognition rate.

TABLE 4

RESULT OF AUDIO VISUAL SPEECH RECOGNITION

	Feature Type		Vector length	Image size	Best Result
	Audio	**Video**			
AV45_128=6+39	MFCC_0_D_A	Blocked DCT	39 audio +6 visual	128x128	76.67% At triphone and mix2
AV44_64=5+39	MFCC_0_D_A	Blocked DCT	39 audio +5 visual	64x64	80% at triphone
AV41_64=2+39	MFCC_0_D_A	Blocked DCT	39 audio +2 visual	64x64	85% At triphone and mix2
AV42_64=3+39	MFCC_0_D_A	Blocked DCT	39 audio +3 visual	64x64	91.67% at Mix2
AV45_64=6+39	MFCC_0_D_A	Blocked DCT +PCA	39 audio +6 visual	64x64	88.33% at triphone

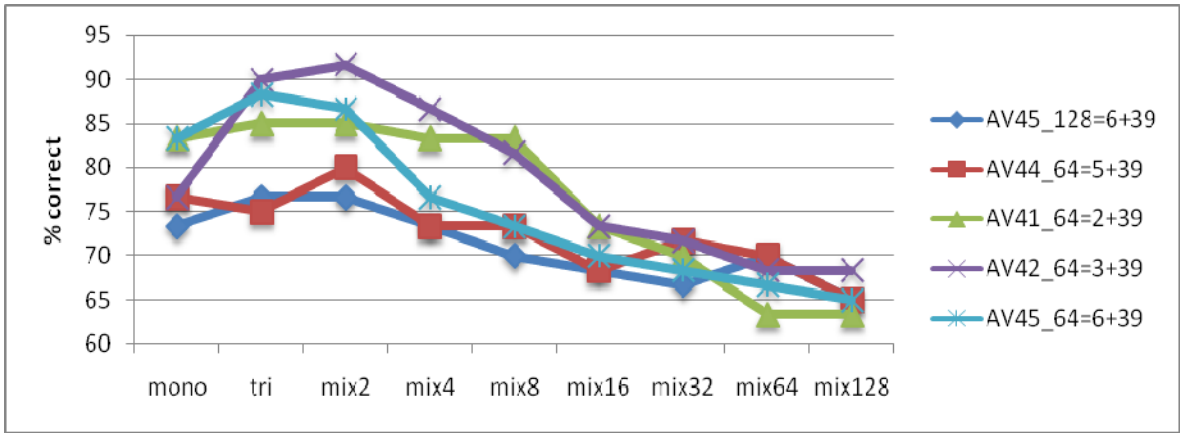


Figure 9: Percentage correct for audio visual speech recognition with different mixtures

- **Experiment 3**: speech recognition using acoustic features only Vs Audio visual feature Vs Video only speech recognition for 100 files

In this experiment, we compare the three feature types to verify the effeteness of the proposed system by obtaining the result for using audio only, video only and the combination of them in audio-visual feature. The results prove that using the audio visual system with blocked DCT visual feature gives better result with 90% recognition rate , in the second stage 88.33% by using audio-visual with blocked DCT and PCA. It means that using visual features with the audio features improve the result with 3.4%. Figure 10 explains the obtained results of getting different visual features and with different sizes for different mixtures.

TABLE 5

RESULT OF THE AUDIO ONLY VS VIDEO ONLY VS AUDIO-VISUAL SYSTEM

	Feature Type		Vector length	Image size	Best Result
	Audio	Video			
A39	MFCC_0_D_A	–	39 audio only	64x64	86.67% At mix2
AV45=39+6	MFCC_0_D_A	DCT	39 audio+6 visual	64x64	90% At triphone
AV45=39+9NewPCA	MFCC_0_D_A	DCT+PCA	39 audio+9 visual	64x64	88.33% At mix2
V6	–	DCT+PCA	6 visual only	64x64	70% At triphone

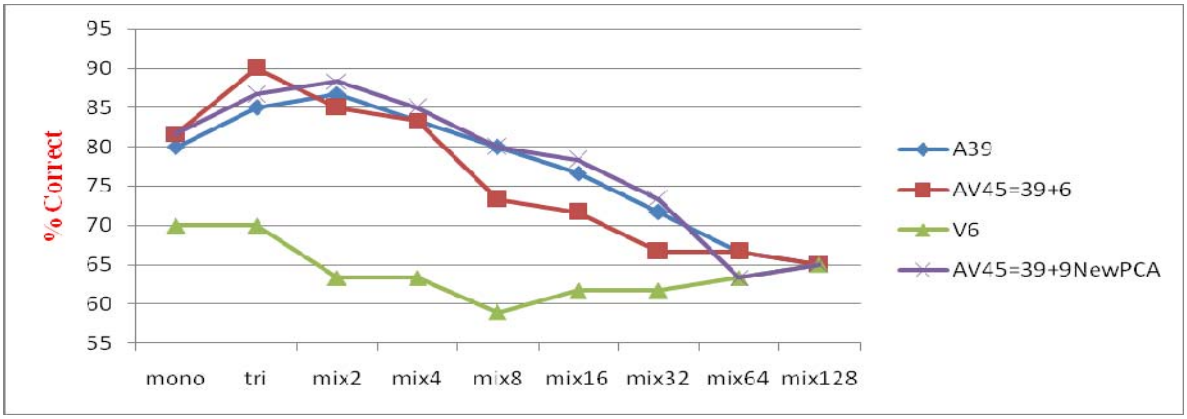


Figure 10: percent correct for Audio only Vs Video Only Vs Audio-visual

- **Experiment 4**: speech recognition using acoustic features only Vs Audio visual feature Vs Video only speech recognition for total database.

From the results obtained we can see that using visual features which are extracted by blocked DCT and PCA with the audio features give enhancement for the performance of the recognition process and give more enhancement in case of noisy system as shown in figure 11.

TABLE 6

RESULT FOR AUDIO ONLY AND DIFFERENT AUDIO VISUAL TECHNIQUES FOR TOTAL GRID DATABASE

A 39	MONO	TRI	MIX2	MIX4	MIX8	MIX16	MIX32	MIX64	MIX128
WORD: %Corr	72.13	92.82	94.01	95.54	96.79	97.73	98.45	98.81	99.25
SENT: %Correct	15.99	64.11	68.79	75.77	82.04	86.81	90.96	93.6	95.51
AV 39+3zigzag									
WORD: %Corr	71.61	91.74	93.35	95.03	96.66	97.83	98.57	99.04	99.35
SENT: %Correct	15.28	59.75	65.99	73.65	81.49	87.7	91.94	94.53	96.22
AV 39+6(5DCTzigzag > 6PCA)									
WORD: %Corr	72.27	91.89	93.29	95.29	96.68	97.71	98.5	99	99.28
SENT: %Correct	16.3	60.82	66.73	75.22	81.89	87.3	91.39	94.16	95.73

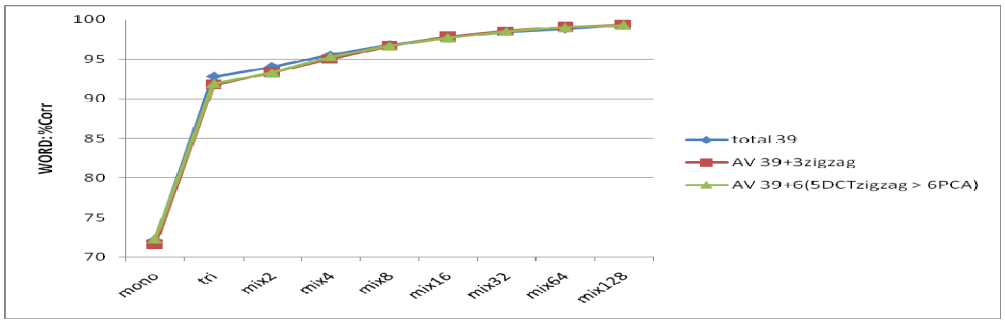


Figure 11: percent correct for Audio only Vs Video Only Vs Audio-visual for total GRID database

4 CONCLUSION

In this paper, we provided a brief overview of the basic techniques for automatic recognition of audio-visual speech, proposed in the literature over the past twenty years, with particular emphasis in the algorithms used in our speech reading system. The two main issues relevant to the design of audio-visual ASR systems are: First, the visual front end that captures visual speech information and, second, the integration (fusion) of audio and visual features into the automatic speech recognizer used. Both are challenging problems, and significant research effort has been directed towards finding appropriate solutions. This study investigates the effect of adding Discrete Cosine Transform Coefficients DCT of mouth region as visual features which dimensionality are reduced by using Principle component analysis PCA with audio features. The proposed system is tested on the standard database, GRID sentence database. Speaker dependent and speaker independent experiments are tested and change DCT visual feature size and using PCA are applied and compared. It was found that adding the whole upper left corner region of DCT coefficients matrix with using PCA can improve the performance of AVASR. From the experiments given in this paper we find that the optimal number of audio vector size is 39 it gives enhancement for the recognition rate by 6.6%. The results indicate that the best recognition is obtained by using 64x64 image size with the blocked DCT and PCA feature extraction with best result obtained when using audio visual feature with size of vector 42 and image size 64x64 which is 91.67% recognition rate. It means that using visual feature with the audio feature improves the result with 5%. Also when testing the system under noisy environment it improves the result.

REFERENCES

[1] A.N. Mishra, Mahesh Chandra, Astik Biswas, and S.N. Sharan, "Hindi phoneme-viseme recognition from continuous speech", *International Journal of Signal and Imaging Systems Engineering (IJSISE)*, Vol. 6, No. 3, pp. 164-171, 2013.

[2] Estellers, Virginia, Thiran, and Jean-Philippe, "Multi-pose lip reading and audio-visual speech recognition", *EURASIP Journal on Advances in Signal Processing*, pp.1-23, 2012.

[3] Burnham, Douglas, et al., eds. Hearing Eye II: The Psychology of Speech reading And Auditory-Visual Speech. Psychology Press, 2013.

[4] Massaro, Dominic W., and Jeffry A. Simpson. Speech perception by ear and eye: A paradigm for psychological inquiry. Psychology Press, 2014.

[5] Stork, David G., and Marcus E. Hennecke, eds. *Speech reading by humans and machines: models, systems, and applications*. Vol. 150. Springer Science & Business Media, 2013.

[6] Dean, David Brendan. "Synchronous HMMs for audio-visual speech processing." (2008).

[7] Chitu, Alin G., and Leon JM Rothkrantz. "Visual speech recognition." *Information Technologies and control, Sofia, Bulgaria*, 3 (2010): 2-9.

[8] Moore, Brian CJ. An introduction to the psychology of hearing. Brill, 2012.

[9] Hansen, John HL, and Taufiq Hasan. "Speaker recognition by machines and humans: a tutorial review." *IEEE Signal Processing Magazine* 32.6 (2015): 74-99.

[10] Gong, Yifan. "Speech recognition in noisy environments: A survey." *Speech communication* 16.3, pp.261-291,1995.

[11] Ross, Lars A., et al. "Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments." Cerebral Cortex 17.5 (2007): 1147-1153.

[12] Iwano, Koji, Satoshi Tamura, and Sadaoki Furui. "Bimodal speech recognition using lip movement measured by optical-flow analysis." *International Workshop on Hands-Free Speech Communication*. 2001.

[13] G. Potamianos, C. Neti, G. Gravier, A. Garg and A.W. Senior, "Recent advances in the automatic recognition of audio-visual speech, "*Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306-1326, 2003.

[14] Davis, Chris, and Jeesun Kim. "Audio-visual speech perception off the top of the head." Cognition 100.3 (2006): B21-B31.

[15] Bailly, Gerard, Pascal Perrier, and Eric Vatikiotis-Bateson. *Audiovisual speech processing*. Cambridge University Press, 2012.

[16] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," *Speech Communication*., vol. 26, pp. 23–43, 1998.

[17] J. Jiang, A. Alwan, P. A. Keating, B. Chaney, E. T. Auer Jr., and L. E. Bernstein, "On the relationship between face movements, tongue movements, and speech acoustics, " *EURASIP J. Appl. Signal Processing*, vol. 2002, pp. 1174–1188, Nov. 2002.

[18] Hennecke, Marcus E., David G. Stork, and K. Venkatesh Prasad. "Visionary speech: Looking ahead to practical speech reading systems." *Speech reading by Humans and Machines*. Springer Berlin Heidelberg, 1996. 331-349.

[19] Salama, Elham S., Reda A. El-Khoribi, and Mahmoud E. Shoman. "Audio-Visual Speech Recognition for People with Speech Disorders." *International Journal of Computer Applications* 96.2 (2014).

[20] Tiwari, Vibha. "MFCC and its applications in speaker recognition." *International Journal on Emerging Technologies* 1.1 (2010): 19-22.

[21] Steve Young, Mark Gales, Xunying Andrew Liu, Phil Woodland, et al." The HTK Book" ,Version 3.41 , Cambridge University Engineering Department, 2006 , http://www.htk.eng.cam.ac.uk

[22] http://www.mathwork.com

[23] Viola, Paul, and Michael Jones. "Rapid object detection using a boosted cascade of simple features." *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*. Vol. 1. IEEE, 2001.

[24] Viola, P., Jones, M. "Robust real-time object detection." *The IEEE Transactions on Computer Vision* 57(2), 137–154 (2004)

[25] Sagheer, Alaa. "Multimodal Arabic Speech Recognition for Human-Robot Interaction Applications." Applied Mathematics & Information Sciences 9.6 (2015): 2885.

[26] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior, "Recent advances in the automatic recognition of audiovisual speech", Proc. IEEE, Vol. 91, No. 9, pp.1306–1326, Sep, 2003.

[27] P. Scanlon and G. Potamianos, "Exploiting lower face symmetry in appearance-based automatic speech reading." *Proc. Works. Audio-Visual Speech Process. (AVSP)*, pp. 79–84, 2005.

[28] Matthews, etc. "Extraction of Visual Features for Lip reading." *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 2, February 2002.

[29] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou. "Audio-visual speech recognition, final workshop report", *Center for Language and Speech Processing*, 2000.

[30] G. Potamianos, H.P. Graf, and E. Cosatto. "An image transform approach for HMM based automatic lip reading." *Proc. Int. Conf. Image Process*., Chicago, pp.173-177, 1998

[31] G. Chiou and J. Hwang. "Lip reading from Color Video." *IEEE Trans. on Image Processing*, 6(8) , pp.1192-1195, August (1997).

[32] P. Duchnowski, etc. "Toward movement-invariant automatic lip-reading and speech recognition."*Proc. Int. Conf. Acoust. Speech Signal Process*., Detroit, pp109-11,1995.

[33] M. Heckmann, etc. "DCT-based video features for audio-visual speech recognition." *Proc. Int. Conf. Spoken Lang. Process. Denver*, USA. September 2002. 9, pp. 1925-1928

[34] A.V. Nefian, etc. "Dynamic Bayesian networks for audio-visual speech recognition." *EURASIP Journal on Advanced Application in. Signal Processing*, Nov. 2002, pp.1274-1288.
[35] G. Potamianos, etc. "A cascade image transform for speaker independent automatic speech reading." *IEEE International Conference on Multimedia and Expo*, 2000, Volume: 2, pp: 1097 -1100.
[36] Gales, Mark, and Steve Young. "The application of hidden Markov models in speech recognition." Foundations and trends in signal processing 1.3 (2008): 195-304.
[37] Cooke, Martin, et al. "An audio-visual corpus for speech perception and automatic speech recognition." *The Journal of the Acoustical Society of America* 120.5, pp. 2421-2424, 2006.
[38] Radha, V., and C. Vimala. "A review on speech recognition challenges and approaches." doaj. org 2.1 (2012): 1-7.

BIOGRAPHY

Amr M. Gody received the B.Sc. M.Sc., and PhD. from the Faculty of Engineering, Cairo University. Egypt, in 1991, 1995 and 1999 respectively. He joined the teaching staff of the Electrical Engineering Department, Faculty of Engineering, Fayoum University, Egypt in 1994. He is author and co-author of about 40 papers in national and international conference proceedings and journals. He is the Acting chief of Electrical Engineering Department, Fayoum University in 2010, 2012, 2013 and 2014. His current research areas of interest include speech processing, speech recognition and speech compression.

Mohamed H. Farouk received the B.Sc. in Electronics Engineering from the Faculty of Engineering, Cairo University, Egypt, in 1982. He received the MSc and PhD. of Engineering Physics from the Faculty of Engineering, Cairo University, Egypt, in1988 and 1994 respectively. He joined the teaching staff of the Electrical Engineering Department, Faculty of Engineering, Cairo University, Egypt in 1984. His Current Position is full Professor, Engineering Math. & Physics Dept., Faculty of Engineering, Cairo Univ. from 2007-till Now. He is author and co-author of about 40 papers in national and international conference proceedings and journals.

Eslam E. El Maghraby received the BSc (Honours) degree in Communication and Electronics from Faculty of Engineering, Fayoum University in 2008. She received the MSc degree in speech recognition systems from faculty of engineering, Fayoum University in 2013. She is currently a PhD student at the Faculty of Engineering-Fayoum University. She is working as Assistant Lecturer at Information System Department at Faculty of Computers and Information, Fayoum University. Her research interest is in signal processing and computer networks.

TRANSLATED ABSTRACT

تحسين جودة ودقة أنظمة التعرف على الكلام باستخدام اشارة الكلام الصوتية والبصرية

اسلام المغربى، * عمرو جودى، ** محمدهشام فاروق*

قسم الهندسة الكهربية- كلية الهندسة- جامعة الفيوم
***قسم الرياضيات و الفيزيقا الهندسية- كلية الهندسة- جامعة القاهرة*
[1]eem00@fayoum.edu.eg
[2]amg00@fayoum.edu.eg
[3]mhesham@eng.cu.edu.eg

ملخص

بالرغم من الجهود المبذولة خلال العقود الماضية للوصول إلي أعلي درجات التعرف علي الأصوات مازالت الأنظمة التي تم الوصول إليها غير دقيقة وغير مناسبة للتطبيقات الحيايتية الحقيقية وخصوصاً تلك التي توجد في أوساط بها الكثير من الضوضاء .معظم أنظمة التعرف علي الأصوات تعتمد علي الاشارة الصوتية كمصدر وحيد للصوت وتقوم بإهمال الجزء البصري المصاحب له. الدمج بين الاشارة الصوتية والبصرية المصاحبة للصوت يقدم وعودا بالوصول لدرجة أعلي للتعرف علي الصوت ودقه أفضل من التي يمكن الحصول عليها من خلال استخدام الاشارة الصوتية فقط. هذا البحث يهدف إلي بناء نظام للتعرف علي الأصوات بإستخدام الاشارة الصوتية بالاضافة الي الاشارة البصرية المصاحبة للصوت لمجموعة من الجمل التي تحتوي علي كلمات منطوقة باللغه الانجليزية. يتم استخراج خصائص الاشارة الصوتية باستخدام خاصية MFCC واستخدام خاصية DCT لاستخراج الخصائص من الصورة المصاحبة للصوت و خاصية PCA استخدمت لغرض تقليل حجم الخصائص المستخرجه من الصورة وذلك لسهولة التعامل معها و للعمل علي سرعة أداء النظام المقترح. يتم دمج الخصائص المستخرجة من الصوت والصورة المصاحبة له لتدريب نظام التعرف علي الأصوات باستخدام HMM للتعرف علي الكلمات المنطوقة عن طريق استخدام اداة HTK. تم اختبار كفاءة النظام المقترح من خلال تطبيقه باستخدام واحدة من أكبر قواعد البيانات للصوت والصورة معا وهي قاعدة بيانات GRID. من خلال تحليل النتائج للنظام المقترح المعتمد علي الصوت والصورة معا نجد انه يقدم كفاءة أعلي ومعدل تعرف اكبر بمقدار 4% عن استخدام الصوت فقط.