Audio Visual Arabic Speech Recognition using KNN Model by Testing different Audio Features

Esra J. Harfash
Department of Computer Science
College of Science
University of Basra, Iraq

Diyar H. Shakir
Department of Computer Science
College of Science
University of Basra, Iraq

ABSTRACT
The most important challenges in AVSR and the focus of most research are the features that are extracted, and when combined give better results. The other challenge is the resulted feature here of nature are large in size, then prefers here to reduce the features by use of an appropriate way to reduce these data with ensure have their properties after downsizing. The System that is presented in this research is for recognition a group of Arabic words voices, from one to ten words. In the acoustic parts the features were extracted of coefficients MFCC, LPC,FFT to be determine which type of these features is efficient in AVSR .All these types of feature are showed efficient results but MFCC is the best. The visual features are calculated of DCT matrix, and the features are extracted by applying the zigzag scan. In the reduction features stage, several methods of data reducing have been implemented; they are LDA, PCA and SVD. Each method are applied to the data separately. The KNN models are used in the stage of recognition, where the testing is implemented on dependent and independent database of words from one to ten. The final results that obtained are efficient and encouraging.

General Terms
Automatic Speech Recognition, Audio Visual Speech Recognition, Mouth detection.

Keywords
Audio-Video Speech Processing, Automatic Speech recognition, Mouth detection, Discrete cosine transformation, Visual Features

1. INTRODUCTION
Speech recognition is utilized in many human-machine communication systems. However, noise that often will appear in real world application environments is still highly challenging for ASR in terms of extracting reliable audio features. One possibility to control this problem is using the visual modality, i. e. significant features of the speaker's face or mouth, to upgrade the recognition results. Many attempts have been made during recent years to incorporate visual information into the process of recognizing speech. For that aim, beneficial speech material, containing both audio and video data that have been recorded in the same instant, is desire to design, evolve and test powerful algorithms for AVSR [29].

Then the AVSR is a multimodal classification problem which is well suited to this analysis, since it involves time-varying signals from two modalities, audio and video, which have different rates and properties. For example, they have different temporal resolutions. The audio file has a sampling rate of tens of kHz, while video has thousands of times less, that is, only tens of temporal samples per second. They also have a different dimensionality, as video has two spatial dimensions and a temporal one, while audio only has a temporal dimension [30].

The utility of the visual modality in human speech is now well understood and plays a very significant role in both perception and production speech [31].

The AVSR is the action that integrating the disciplines of image processing, visual/speech recognition and multi-modal data integration [4], where the visual information is the complemented to acoustic information in human speech perception, especially in noisy environments. Where the Humans can disambiguate an acoustically confusable phoneme exploitation visual information since numerous phonemes which are closely related to each other acoustically are very dissimilar from each other visually [32].
Visual information such as signs, expressions, head-position, eyebrows, eyes, ears, mouth, teeth, tongue, cheeks, jaw, neck, and hair, could enhance the implementation of machine speech recognition [33, 34].

The presented work in this research introduces system of audio-visual automatic speech recognition for Arabic words recognition. In particular, the audio signal features with the features of the speaker's mouth movement are used to provide efficient features for this system. Different features and methods are adopted here to get the high performance in audio visual Arabic speech recognition.

2. AUDIO VISUAL SYSTEM
The aim of speech recognition systems is to determine spoken language correctly from features delineating the speech production process. The process of AV-ASR will be executed on records videos that have been recorded by several speakers; firstly the video must be partitioned in two parts the audio part and visual part after that each of them will be processed separately to gain the interesting features for the recognition. Figure 1 show the general form of AVSR system [26].

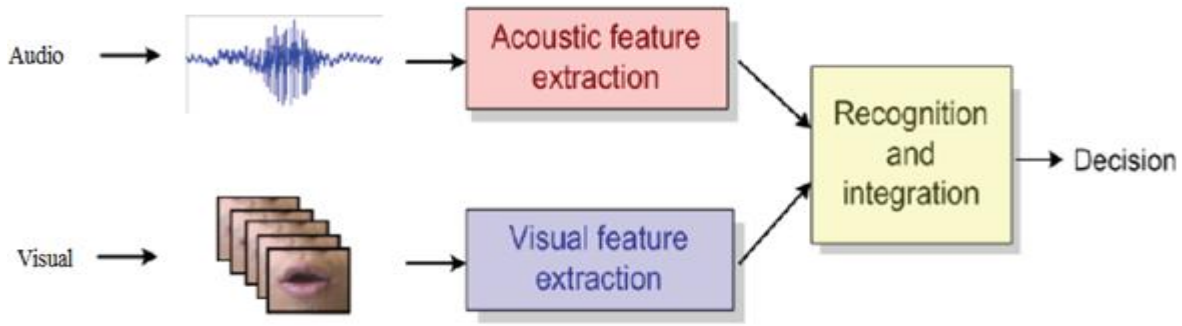

Fig 1: The general Organization of AVSR.

The distant between camera and speaker is kept constant in order to get seemly visual utterance. When the input is acquired, the acoustic and visual feature extraction will be preprocessed separately and further used for recognition and integration of utterance.

The AV-ASR systems consist of three main parts [35]:

1. a visual frontend (i.e., mouth detection and tracking and visual characteristics extractor),
2. Audio front end (audio preprocessing and features extraction)
3. an audiovisual fusion strategy, and
4. speech recognition.

3. RELATED WORKS

Audio VSR Systems are researched through a big deal of studies, some of them are mentioned below:

- In 2012, Galatas. G. et.al .disclosed a study that used the facial data for a multimodal speech recognition system as an additional information, snatched by the Kinect, for the purpose of supporting ASR performance and robustness to noise. Depending on appearance based features for extracted visual features from mouth region-of-interest (ROI), gained from the discrete cosine transformation DCT, this process also used the linear discriminated analysis LDA for further act. And a two-stage LDA was applied to the visual features. The Viola-Jones algorithm has been used for "Face and Mouth regions" detecting. The BAVCD database has been adopted [10].

- In 2014, Silber-Varod. V. et.al., proposed a novice viewpoint of the challenges of "ASR": rather than redusing "word error rates (WER)", center on keyword recognition. There data set focused on Hebrew language, which have not yet approached acceptable an adequate level of recording. A forty-minute recording set, which includes audio books and academic lectures, was used for examining the performance of Hebrew ASR systems, and comparing it to stenographer recordings of the video lectures, whilst shedding light on keyword recognition [11].

- In 2003,Potamianos.G. et.al., two different environments that give important challenges to robust visual processing were presented an investigated audio ASR: (a) Typical offices, the data have been recorded by a movable PC with an inexpensive web camera, and (b) cars, with data collected at three approximate speeds. They made use of MFCC for audio features extraction, and DCT and LDA for visual features extraction. Then, the act of this system is reported on these two sets of database and benchmarked against "visually clean" data recorded in a studio-like environment [12].

- In 2008, Helge Reikera H. et.al., worked hard to cause existence of a basic AVASR system that used MFCCs as acoustic features, and effective appearance model (AAM) parameters as visual features, and for modeling the distribution of audio-visual speech feature they presented Gaussian Mixture Model(GMM). The Principle Component Analysis (PCA) and LDA have been used for minimizing the dimensions of the extracted features to lower dimensions. Using the CUAVE database, The AVASR system was implemented and tested [13].

4. THE PRESENTED WORK

This AVSR system for recognizing some Arabic sounds for isolated words. The words sounds that have been used in this system are words sound of numbers (1-10).
The input here consists of two files the Audio file and the visual file for one video, each input file is processed separately until the combination step where the video features combined with the sound features:

1. AUDIO FILE: The acoustic parts of spoken word file read completely, here the audio file as in figure 2 will be framing during subsequent work stages:
   X=AUDIO FILE;

2. VISUAL FILE :Each visual input read as a sequence of frames (sequence image) ,and each frame (image) processed separately:

   MOV = VISUAL FILE;
   Where: MOV=$VFram_1$, $VFram_2$,……, $VFram_N$. ,and $VFram_i$ is one frame of visual input of video and i=1…N .

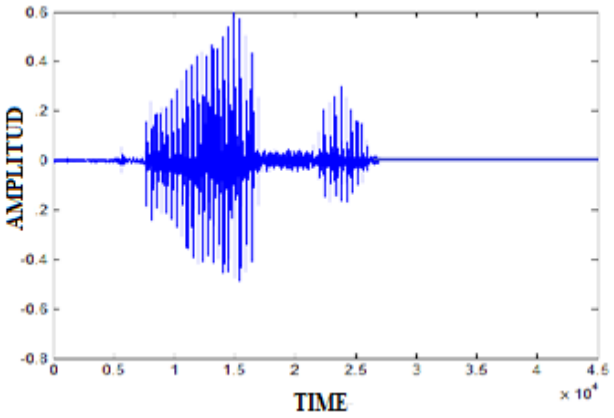

Fig 2: The audio parts of word (one)

4.1 Audio-front end

The following steps are the pre-processing that applied on sound file.

A. In this work the normalize is applied as following :

$$\text{Norm}(X) = \left(\frac{\sum_i^N X_i}{N}\right) / \max|X| \quad \ldots(1)$$

Where X is the audio part which is framed later.

B. The audio signal is framing into short times with 256 samples for each ferame ,then the resulted sequence of frames are : $frame_1$,$frame_2$…, $frame_N$

Where z is the resulted number of frames for the input word signal.

C. The Hamming window is used here :
$WFrame_i = Frame_i * hamming(length(Frame_i))$

4.2 Visual front-end

In this work begins with identification the face of the speaker, and in the next step tracking the mouth and extracting the region of interest (RIO), the following steps are the pre-processing that applied on Visual file.

A. The Detected mouth algorithm has been applied on each frame in the video where: detectmouth ($Vframe_i$)= ROI_i where i=1 to N , and ROI_i is the image of mouth that detected from one frame($Vframe_i$) of input video:

1- Read the image(one frame I_j),where I[X1:Y1 ,X2:Y2]
2- Detecte the face region of I.
3- Extract the face region .
4- Detecte the mouthe region of face ,where Mouth=face[X1:Y1 , X2:Y2+2Y1/3]
5- Extract mouth region RIO_j
6- Increment j by one , if j less than or equal 25 go to step 1 else End.

For this algorithm can see the following:

- In this technique of mouth detection, the face is determined first and its limits after its detection are X_1 and Y_1 are the width, height, X_2 and Y_2 are the upper left corner of face region.
- After specify the face , then the area of the mouth is tracked in the face box region, this area is in the last third of the face region, and has X_1 and $Y_1/3$ are the width and

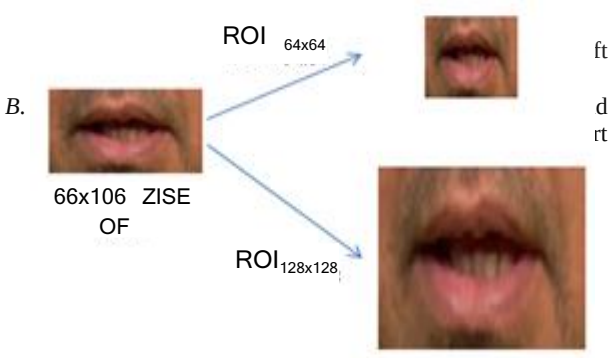

Fig 3: Apply two type of resize on ROI.

4.3 Feature extraction

This stage is the feature extraction that consists of the analysis Audio and Visual parts of video . The features that are relied in this research are: the DCT that applied on visual file, and the MFCC, LPC and FFT are applied on audio file. This stage is now discussed in more detail.

4.3.1 Audio Feature Extraction

In this stage multiple Audio are extracted and tested: the features are extracted from MFCC , LPC and FFT coefficient as following:

Mel Frequency Cepstral Coefficients (MFCC): The Following are the steps that applied to calculating the MFCC:

1- Calculate MFCC coefficients : MF_i= MFCC($Wfram_i$), i=1 to N ,Where MF_i is the coefficients of MFCC for each frame $Wfram_i$.

2- Keep the 13 MFCC coefficients. This mean :

MFCC (Audio)= MF_1, MF_2,……, MF_{25} , where each MF_i is 1..13 coefficients

Linear Predictive Coefficients (LPC): LPC is one of the most powerful speech analysis techniques that provide extremely accurate estimates of speech parameters. Then we are used it as features here. The following is the steps of calculating the LPC:

1- Calculate LPC coefficients: L_i =LPC($Wfram_i$) , i=1 to N
Where L_i is the coefficients of LPC for one frame $Wfram_i$

2- Select 12 value of LPC coefficients of L_i :
LPC (Audio)=L_1,L_2,……,L_{25} , where each L_i is 1..12 coefficients.

Fast Fourier Transform (FFT):To representing the given audio signal in frequency domain is done via Fast Fourier Transform (FFT) which implements as following :

1- Calculate DFT coefficients for each frame by Appling FFT algorithm :F_i =FFT($Wfram_i$) ,Where i=1 to 25 , F_i has 1:N/2 coefficients, and N is the length of each frame ,where N=256

2- Find magnitude spectrum and maximum energy in each band :
$M_i = \max(|F_i(n1:n2)|)$ …..(2)

Where i=1 to 25 ,n1 and n2 begin with value 1 and 16 respectively.And continue n1 and n2 increments each time by 16 to reach n1 to 128 and here stops the calculate of M_i values .The resulted length of M_i is 8 value.

3- The Implementation of Equation (6) continue until i>N.
Figure 4 show the original speech signal and its MFCC , LPC and FFT coefficients.

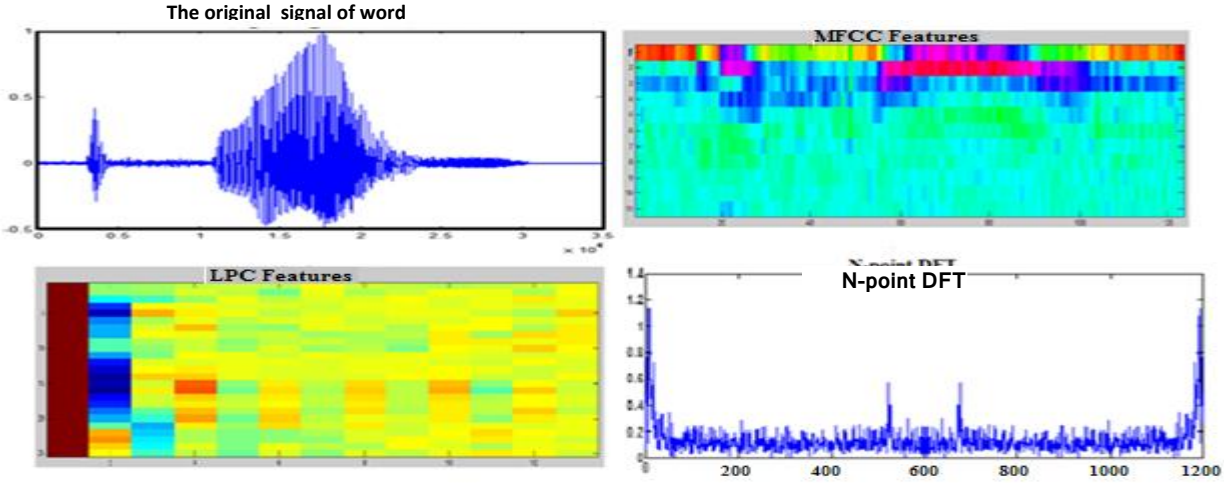

Fig 4: show the original speech signal and its MFCC , LPC and FFT coefficients

4.3.2 Visual Features Extraction

The Visual feature extraction process applied on the (ROI) in the size (64 x 64) and (128 x 128) separately. And the main method that we are used to extract the features is the DCT.

The DCT to be calculated, the images must be square and gray scale (these operation is implemented in previous operations: Di=DCT(ROIi)

Figure 5 show an example to apply DCT transformation on mouth (RIO) that represents one frame.

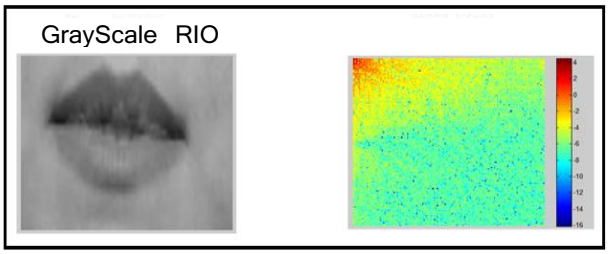

Fig 5: The ROI and its corresponding DCT values

The features that are adopted in this work using DCT coefficients are selected as following:
Selected the upper 40x40 matrix of DCT and then Applied the zigzag scan on this upper 40x40 matrix then select the first 50 value through the experiences and stored.

4.4 Dimension Reduction

The process of reduction data's dimension is mean that transform the data space from high dimensional (HD) to low dimensional (LD), since many problems like recognition systems can be done in more accurately with the reducing data. This process should be done carefully to maintain all the important features of our data.
Different reduction methods are used in this work to reduce the features dimensions (in order to test the efficiency of each method in reducing data and get good recognition results) .The DR carried on both audio signal and visual signal, The following is an explanation of the use of these methods in this research.

A- Linear Discriminate Analysis (LDA): LDA is an important method in problem of recognition, since it finds the linear combination of features in order to find the separation of two or more classes. The following is the requirement of LDA to be implemented:

1. For audio or visual features of size ($N_X M_X Z$), where Z is the number of input files, is converted to a two-dimensional vector because the LDA method works with two dimension matrices,reshaping size of this features.
2. The LDA requires a vector named group, where through which the classes are determined here and the length of it must be equal to Z length.

B- Principle Component Analysis (PCA) : PCA is linear transformation methods that reduce the data space to become have maximum variance. The PCA method works with two dimension matrix, then for features of size ($N_X M_X Z$).

For example, for any features of size $N_X M$ the result after PCA implementation is reduced features with size $N_X K$, where K is the eigenvector that corresponding to the maximum value of eigenvalue.

C- Singular Value Decomposition (SVD)We were used this technique, because it is relatively straightforward and have the ability of dimensionally (number of column of features) reduce:

[Singular vectors] = SVD (features $_{N x M}$).
The singular vectors $_{(N x M)}$ were ordered descending according to the order of singular value matrix $_{(N x M)}$, and then selected the first d ordered vectors.

4.5 Features Integration

The Feature integration is an essential process for the AVSR, where the audio features merging with visual features. The following is the description of combination process:
(If audio features with long 1,...,M_1 features and visual with long 1,...M_2 features then after the fusion the AV result is 1,…(M_1 +M_2) features.)
The fusions are done in this work is Early integration (EI), and we are implemented in two ways:

4.6 Recognition

At this stage, the KNN model has been used for testing the data. In the problem of recognition, the input to the nearest neighbors algorithm (K-NN) is consist of the M training classes in the feature space, and the output depends on the K-NN classification, where the output is a class membership. In following steps the description of classifies of input test word into one of the groups in Training set using the nearest-neighbor method.
From the above step ,we see that:

- group array must be for training and testing. each element in group array defines the group to which the corresponding row of training set belongs to it)
- Through implement the KNN, the distances from all points in the training set A to all points in the test set B are calculated.
- The final result CLASS represent the type of classes that B belongs to it. CLASS indicates which group each row of tested w has been assigned to, and is of the same type as Group.

5. PERFORMANCE EVALUATION

In this system there are various data that extract from the Audio and Visual parts**,** all these data have been tested according to the following distributed:

1- MFCC +DCT
2- LPC +DCT
3- FFT +DCT
4- In each stat of above 1, 2 and 3 the DCT features are applied with 64x64 and 128 x128 mouth sizes separately.

The results that are obtained with KNN model of this database are shown in table 1 for independent database consist of 8 speaker and Table 2 for dependent database. Table 1 and Table 2 show the accuracy recognition with different audio features that obtained on each of 64x64 and 128x128 images sizes.

Table 1: Accuracy of KNN of no. independent database

Audio +visual features	PCA	SVD	LDA
MFCC+DCT$_{64X64}$	73.56481	61.75925	75.13889
MFCC+DCT$_{128x128}$	73.56481	61.71296	69.16667
LPC+DCT$_{64X64}$	68.47222	59.86111	57.63889
LPC+DCT$_{128X128}$	68.33333	60.13888	55.64815
FFT+DCT$_{64X64}$	68.79629	51.85185	56.57407
FFT+DCT$_{128X128}$	68.93518	52.08333	55.41666

Table 2: Accuracy of KNN of NO. dependent database

Audio +visual features	PCA	SVD	LDA
MFCC+DCT$_{64X64}$	79.33333	55.33333	96
MFCC+DCT$_{128x128}$	78.66666	53.77777	94.66666
LPC+DCT$_{64X64}$	68.44444	41.55555	81.55555
LPC+DCT$_{128X128}$	68.66666	40.22222	79.77777
FFT+DCT$_{64X64}$	47.33333	32.44444	83.77777
FFT+DCT$_{128X128}$	49.11111	33.11111	80.66666

The figure 6 shows the char measure of the differences in accuracy by applying different audio features, and chart 7 show the average accuracy of reduction methods that are used in this system with independent database.

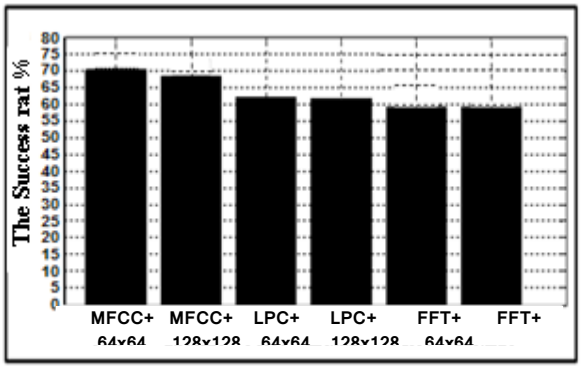

Fig 6: Accuracy recognition of different Audio features of NO. Database of independent system by KNN model

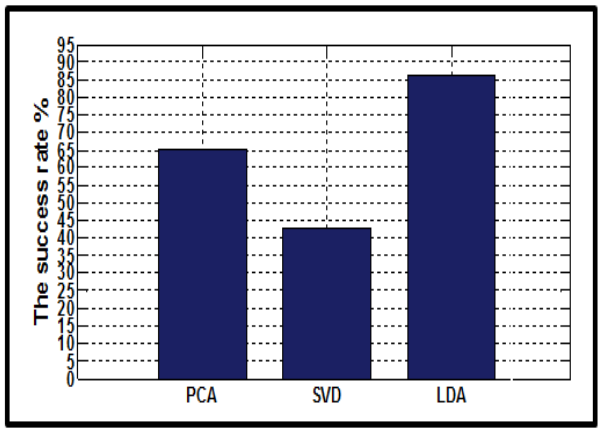

Fig 7: The accuracy differences recognition among the reduction methods

Also the charts in figure 8 and figure 9 show the results that are obtained by implement different audio feature of dependent database with different reduction methods and can be concluded that All audio features are gave good accuracy results , but the MFCC had better results than other audio features .

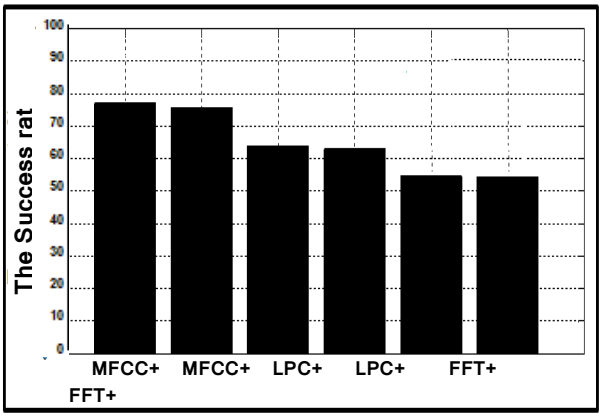

Fig 8: Accuracy recognition of different Audio features

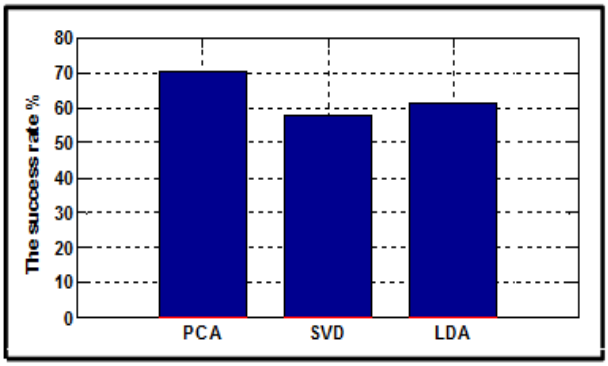

Fig. 9: The difference in accuracy recognition among the reduction methods in dependent system

6. CONCLUSIONS AND FUTURE WORK

There are some conclusions have been drawn based on the presented research as following:

1. In adoption of 64x64 and 128x128 (spatially 64x64) frame dimensions is represent appropriate in aspect of size ,where the time of processing is reduced ,and also is leaded to the desired results.
2. The results with MFCC , LPC and FFT is suited and can be depended in AVSR system after combining them with visual features, but the MFCC audio features in our work are considered very successful features that are get better results
3. The DCT features that extract by zigzag , appeared the good recognition results
4. This research shows that LDA and PCA methods are efficient ways in select and reduce the data to ensure the best results. The accuracy with SVD are generally less than the results with LDA and PCA. But the SVD method is good method in reduction space spatially with the small classes' database.
5. The KNN are proved their ability to classify and recognize the test words sound in AVSR system.

For more accuracy recognition can be combining more than one type of feature, such as MFCC+LPC feature in Audio parts, DCT+WAVLET feature in visual to increase the accuracy recognition. And also can be Using more than one recognition method at the stage of recognition and the result of each one is combing to giving the final decision.Finally, Appling another types of features that extract from images of mouth ,such as the SURF algorithm to find blob features, HOG features that re-encode local shape information from regions within an image, and another technique.

7. REFERENCES

[1] Vorwerk A., Wang X., Kolossa D., Zeiler S., and Orglmeister R., "WAPUSK20 – A database for robust audiovisual speech recognition", Chair of Electronics and Medical Signal Processing , EMSP, University of Berlin, Einsteinufer 17, 10587 Berlin, 2011.

[2] Potamianos G., Neti C., Luettin J., and Matthews I., "Audio-visual automatic speech recognition: an overview". Issues in audio-visual speech processing. MIT Press, 2004.

[3] Lucey S., Chen T., Sirdharan S., and Chardran V.," Integration Strategies for Audio-visual Speech Processing: Applied to Text Dependent Speaker Recognition", Queensland University of Technology, Australia, 2004.

[4] Pao T.L., and Liao W.Y., "AVSR for Testing AV Database", Department of Computer Science and Engineering, University of Tatung, Taipei, Taiwan, R.O.C, 2006.

[5] Kratt J., Metze F., Stiefelhagen R., and Waibel A.," Large Vocabulary Audio-Visual Speech Recognition Using the Janus Speech Recognition Toolkit", Interactive Systems Laboratories University of Karlsruhe , Germany, 2004.

[6] Potamianos G., Neti C., and Deligne S., " Joint Audio Visual Speech Processing for Recognition and Enhancement". Proceedings of AVSP, 2003.

[7] Goecke R., and Potamianos G., " Neti. Noisy Audio Feature Enhancement using Audio-Visual Speech Data". ICASSP 02, 2002.

[8] Bord P., Varp A., Manz R., and Yannawar P., "Recognition of Isolated Words using Zernike and MFCC features for AVSR", Department of Science and Technology (DST), India, 2011.

[9] Gagnon L., S., Foucher F. L., and Boulianne G., "A simplified audiovisual fusion model with application to large-vocabulary recognition of French Canadian speech", CAN.J.ELECT. COMPUT. ENG., VOL. 33, NO. 2, SPRING 2008.

[10] Galatas G, Potamianos G., and Makedon F., "AVSR Incorporating Facial Depth Information Captured by the Kinect", 20th European Signal Processing Conference EUSIPCO, Bucharest, Romania, August 2012.

[11] Silber-Varod V, and Geri N., "Can ASR be Satisficing for Audio/Visual Search? Keyword-Focused Analysis of Hebrew Automatic and Manual Transcription",Online Journal of Applied Knowledge Management, Vol. 2, Issue 1, 2014.

[12] Potamiano G., and Neti Ch., "AVSR In Challenging Environment", Processing of the European Conference on Speech Communication and Technology (EUROSPEECH), PP. 1293-1296, Geneva, Switzerland, sept. 2003.

[13] Reikeras H., Engelbrecht H., Herbst B., and Preez J.D., "AVSR using SciPy", University of Stellenbosch, http://www.SciPy.org/, 2008.