Speech Recognition Using Historian Multimodal Approach

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

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

[1]eem00@fayoum.edu.eg

[2]amg00@fayoum.edu.eg

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

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

Abstract: This paper proposes an Audio-Visual Speech Recognition (AVSR) model using both audio and visual speech information to improve recognition accuracy in a clean and noisy environment. Mel frequency cepstral coefficient (MFCC) and Discrete Cosine Transform (DCT) are used to extract the effective features from audio and visual speech signal respectively. The Classification process is performed on the combined feature vector by using one of main Deep Neural Network (DNN) architecture, Bidirectional Long-Short Term Memory (BiLSTM), in contrast to the traditional Hidden Markov Models (HMMs). The effectiveness of the proposed model is demonstrated on a multi-speakers AVSR benchmark dataset named GRID. The experimental results show that the early integration between audio and visual features achieved an obvious enhancement in the recognition accuracy and prove that BiLSTM is the most effective classification technique when compared to HMM. The obtained results when using integrated audio-visual features achieved highest recognition accuracy of 99.07%, this result demonstrates an enhancement of up to 9.28% over audio-only recognition for clean data. While for noisy data, the highest recognition accuracy for integrated audio-visual features is 98.47% with enhancement up to 12.05% over audio-only. The main reason for BiLSTM effectiveness is it takes into account the sequential characteristics of the speech signal. The obtained results show the performance enhancement compared to previously obtained highest audio visual recognition accuracies on GRID, and prove the robustness of our AVSR model (BiLSTM-AVSR).

Keywords: ***DCT, MFCC, HMM, BiLSTM, and GRID.***

1 INTRODUCTION

Speech understanding for human is performed by using audio and visual information e.g. movements of speaker lips and tongue, where using lip movements to identify the spoken words is known as Lipreading. Lipreading can be used in addition to audio signal to enhance speech recognition performance for hearing-impaired listeners, and for the case of whisper speech where the performance of audio only speech recognition systems decreases [1]. It also can be useful for people with normal hearing especially in noisy environments [2], [3] as the visual speech signal not influenced by the acoustic noise.

Speech recognition system consists of two main parts, feature extraction and classification process. Choosing the most effective feature extraction method and the best classification technique have been an attractive research topic for decades, with the inventive work introduced by Petajan [4]. Generally, visual feature extraction methods can be classified into three classes: 1) “Appearance or pixel based” which based on a pre-defined region of interest ROI of the lip region and supposes that the whole lip region is informative to speech recognition. It depends on a traditional image compression technique e.g. Discrete Cosine Transforms (DCT) [2], Discrete Wavelet Transform (DWT), Principal Components Analysis (PCA) [5] , and linear discriminate analysis (LDA) [6]. Among these different methods, DCT has been proven to perform equally good or superior to others [7]. This method achieved high accuracy for visual-only speech recognition task because it gives a close representation of the mouth region. Even that appearance-based features are preferred because they do not need restricted lip shape models or hand-labeled data for training, they are vulnerable to the changes in lighting conditions, translations, or rotations of input images, Deep learning can be used to overcome these weaknesses [8]. 2) “Shape or lip contour based”, where a prior template or model is used to describe the mouth area, it faced an information loss [6] because it uses only the width and the height not the whole region of the speaker's lips, example for this method the system introduced by Chowdhary [9] where a scale-invariant feature extraction and shape-index depiction method is used to form robust object recognition system . 3) The combination of 1 and 2 which takes width and height in addition to pixel values of the ROI, example of this method the system introduced by Chan [10] which proposed a visual feature representation combined both geometric and pixel-based features to perform visual-only and audio-visual speech recognition.

Adding the visual speech features to the ASR system is a good choice for speech recognition enhancement, McGurk effect [11] explained the relation between audio and visual features and proved that adding the visual feature to audio ones can impressively change the decision of the recognition process. The significant improvement of visual information to speech recognition for noisy environment encourages researchers to use vision in addition to hearing in a speech recognition system. Potamianos et al. [12] summarized the main components required to build a robust AVSR system.

Choosing the best classification techniques has been a significant research topic for decades either for visual-only or audio-visual speech recognition system; it used to represent the temporal evolution of the speech features, with inventive work introduced by Petajan [4]. Previously HMM was the state of art classification technique for speech recognition system for normal [13] and disorder people [14]. Although, HMM is easier to understand and implement, deep learning proved to be a strong competitor to the HMM classifier and one of the most promising solutions for both audio and visual speech recognition. Deep learning is preferred over HMM due to its robust self-learning mechanism and confirmed performance for speech recognition applications [15], It would probably take more computational time, but the results can be more reliable. The accuracy in certain image recognition and language processing problems is superior when using deep learning [16].

This paper proposes audio-visual speech recognition system by extracting the visual feature in addition to the acoustic features from the speech signal; the classification process is performed by using one of the major Deep Neural Network (DNN) architectures BiLSTM while applying HMM at the same time to compare the obtained results. The proposed model is tested in GRID database to ensure its availability for different size dataset.

A. Problem statement

Audio features are still the main involvement which plays the most important role in speech recognition than visual features. However, in some cases, extract valuable information from the audio only signal is a hard task, such as detecting a person speech from a distance, or understanding a person speaking among a very noisy crowd of people, in these cases, the performance of audio speech recognition is very limited. The tasks of effective visual feature extraction, the successful integration of the acoustic and visual speech modalities, and selecting the most effective classification techniques, have up till now to be successfully addressed.

This paper aims to propose robust and reliable approach to AVSR system by using deep learning classification engine BiLSTM and compared its results to traditional HMM classifier to insure the robustness of the proposed speech recognition system.

B. Paper structure

The remainder of this paper is structured as follows: Section 2 summarizes literature reviews on speech recognition system, and section 3 introduces the algorithms and the main stages for building our model including audio and video feature extraction and classification techniques. The functionality test of the proposed model and their results are shown in Section 4. The obtained results are discussed, and conclusions are made in Section 5.

2 LITERATURE REVIEW

In this section we discuss some of the latest and most relevant algorithms in building speech recognition system which can be divided into two main stage feature extraction and classification process.

In addition to the previously mentioned well-known visual feature extraction methods a lot of researches make a combination of two or more of them to form a new feature extraction method. Acharjya [17] gave a combination of singular value decomposition (SVA) with PCA and Chowdhary [18] illustrated the potential of using PCA with ICA to extract the speech feature from the visual signal. Chowdhary [19] used PCA with LDA in the third stage of their 3D object recognition system, to perform dimension reduction process in order to concentrate on the most effective among a large space of the obtained features. Deep learning can be used to extract the visual feature as done by Noda k [8] where convolutional neural network (CNN) is used to extract the visual features to recognize phonemes the small part of sound [8] and Petridiset al. [20] used LSTM on the extracted Deep Bottleneck Features (DBF) in combination with Discrete Cosine Transform (DCT) features, this approach achieves enhancement of speech recognition up to 5% over DCT. Not all obtained features belong to visual tracking, there redundant features caused in performance degradation, selecting the most effective features from the conventional features is still a challenge for machine learning algorithms. Zhang [21] introduced an approach of adaptive weights-objective function to select the appropriate feature for machine learning algorithms.

Classification techniques can be done by several techniques while HMM is considered to be the most usable classification technique in speech recognition system. Noda k [8] used HMM to recognize phonemes and Koller [22] used it to recognize visemes the visual equivalent of the phonemes, also Goldschen [23] and Tao [1] used HMM in their Lipreading system. Support Vector Machines (SVMs) used by Barnard [3] while H. Ninomiya [24] used DBF to encode input images, G. Potamianos [25] used DCT, and K. Yamaguchi [26] used CNN; all of them used these features with HMMs to classify spoken digits or isolated words. Elham S. [14] introduced an AVSR system for people with dysarthria speech disorder, MFCC is used to extract the acoustic speech signal, DCT coefficients are extracted from the mouth region and concatenate the features vector from both components then applied the HMM classifier. Deep learning achieved promising results in classification stage for speech recognition process. Mrouehet al. [27] employed feed-forward Deep Neural Networks (DNNs) to implement phoneme classification using a nonpublic audio-visual dataset. Petridis et al. [28] completed the previous work given in [29] to obtain speech feature from image pixels and waveform, these features concatenated, at the end of the system bidirectional recurrent network is used to get the final word label. Feng W [30] introduced a multimodal Recurrent Neural Network (RNN) model to consider the sequential properties of audio and visual modalities for AVSR, the audio modality is modeled by using LSTM, and the visual modality is modeled by using CNN plus LSTM RNN, at the fusion part both models are combined by a multimodal layer, the performance of this model is validated on AVletters dataset. Ephrat [31] introduced CNN based end-to-end model to produce an acoustic speech signal using the speaker's silent video frames, this model shows great results for recognizing out-of-vocabulary (OOV) words; which performed on speaker four (S4, female) from GRID [32] dataset. Praveen [33] developed a spoken language processing system, which combined a software BiLSTM-based cell speech recognizer and a hardware LSTM-based language processor to perform Natural Language Processing (NLP) system. A hybrid BiLSTM-HMM and unidirectional LSTM-HMM system is introduced by Alex [34] which proved to improve the phoneme recognition performance.

Choosing large database that have many and variety of speakers is a challenging point in evaluating the performance of AVSR system. GRID is considered to be one of the largest audio visual speech recognition dataset, which consists of complete sentences of continuous English voice commands. There are a lot of researches using GRID database to evaluate the performance of the proposed model but these researches either focus in specific speaker or perform phoneme recognition. Wand et al. [35], and Chung et al. [36] used GRID dataset to form word and sentence-level classification based on fully LSTM architecture, which achieves higher results compared to traditional methods on the same dataset. Shankar M [37] proposed a training algorithm for an AVSR system using deep RNN which is evaluated on GRID corpus and provided a comparison of feature fusion and decision fusion. Recently, Assaelet al. [38] performed labeling by using CNN, LSTM and Connectionist Temporal Classification (CTC) [39] which reports a strong speaker-independent performance on the constrained grammar and the 51 words vocabulary of GRID dataset. Eslam E. [13] proposed an AVSR model that can recognize complete sentences, MFCC is used for audio feature extraction and DCT is used for visual feature extraction from the detected mouth region of the speaker. Principal Component Analysis (PCA) is used to reduce the overall dimension of the combined features vector from audio and visual parts before feeding them to the HMM classifier. There also a lot of researches used Grid to either evaluate their system video only or audio-visual speech recognition system [31] [40] [41] [42].

Expanding on ideas from recent achievements in deep learning researches, Deep learning is used for the classification stage for the proposed system in this paper.

3 PROPOSED MODEL

The proposed deep learning audio-visual speech recognition system is shown in Figure 1. The performance of the proposed can be divided into three stages: Data preparation stage, visual front-end and audio front-end. In the Data preparation stage firstly, we need to extract a synchronous audio file from its corresponding video file and segmenting them to word boundary when using Grid database. In audio and visual front-end, preprocessing and feature extraction operations are performed separately. The audio-visual features integration process is performed in the obtained features from audio and visual signal. Finally, the classification process is performed either by using HMM or Bidirectional LSTM (BiLSTM) to obtain the recognized words.

A. Data preparation stage

Data preparation steps are described in this subsection where steps are performed on audio and video files to prepare them for feature extraction step.

1) *Extract alignment Audio from video file:* using ffmpeg command to extract mono channel audio file from the corresponding video file [13], to obtain audio file which has the same time duration as the corresponding video file.

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

2) *Word boundary segmentation:* In order to perform isolated word recognition, the input video is firstly segmented into short clips which have either isolated words or smaller parts representing phonemes or visemes [43]. The input video file is segmented into isolated words boundary for audio and video files where each video in GRID dataset contains a complete sentence like "bin blue at e 9 now". The frame level alignments file distributed with the dataset were used to get word level segmentations of the video in the aid of MATLAB [44] program. This segmentation caused the training dataset to consist of 6 words in each sentence, for the 1000 sentences so we have 6000 single words per speaker (each speaker has 1000 sentences). Table I gives an example for the alignment file for bbae9n.mpg file and the output of segmenting the video and audio to the word boundary according to the corresponding alignment file. The output of SFS program [45] explains the word boundary for bbae9n.wav file as shown in figure 2.

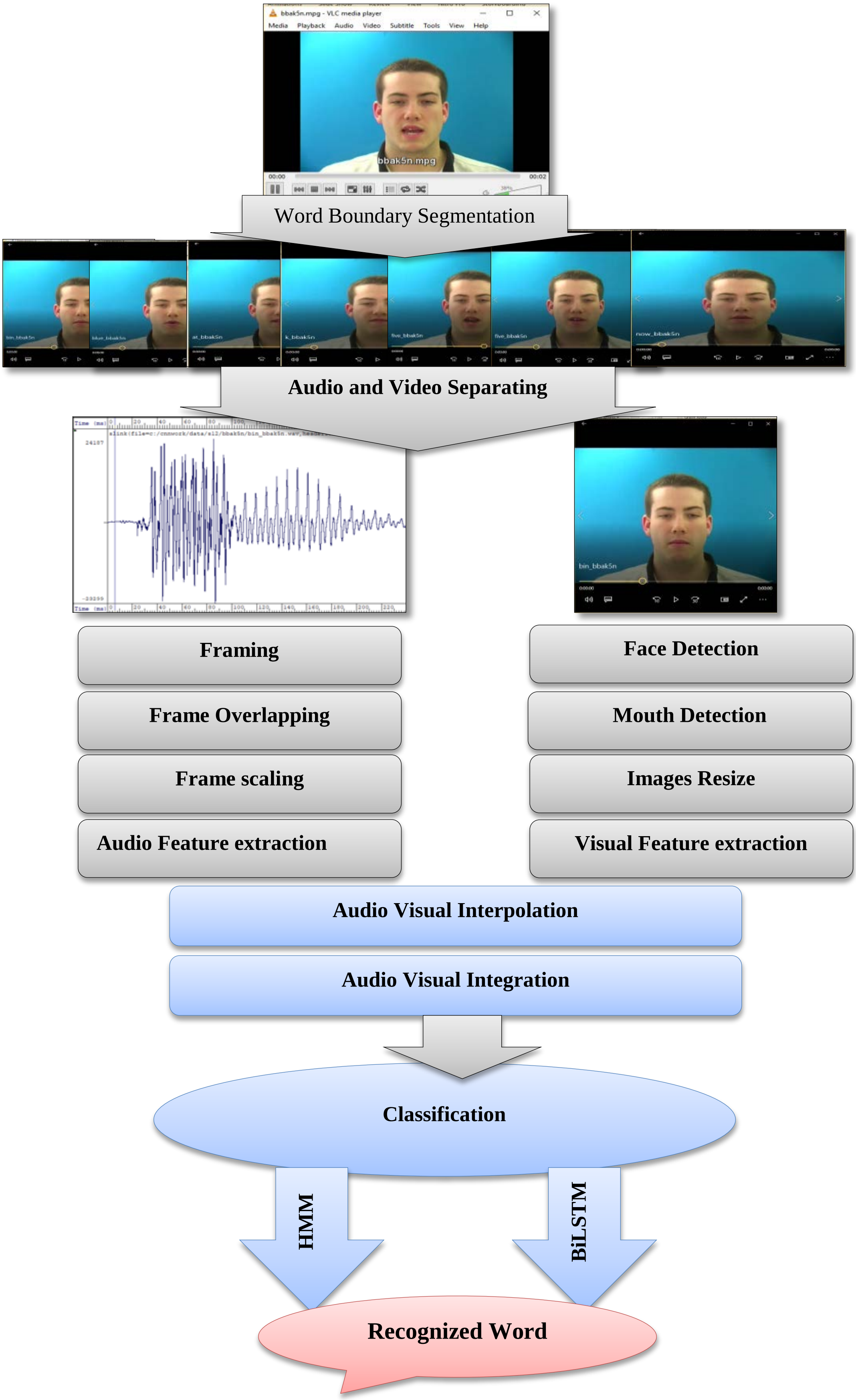

Figure 1: Structure of the proposed AVSR model

TABLE I

SEGMENTED WORD BOUNDARY AUDIO AND VIDEO FILES ACCORDING TO THE CORRESPONDING ALIGNMENT FILE FOR VIDEO FILE BBAE9N.MPG

Audio	Video	Alignment file
bin_bbae9n.wav	bin_bbae9n.avi	0 13500 sil 13500 20000 bin 20000 25250 blue 25250 28000 at 28000 31500 e 31500 42250 nine 42250 53250 now 53250 74500 sil
blue_bbae9n.wav	blue_bbae9n.avi	
at_bbae9n.wav	at_bbae9n.avi	
e_bbae9n.wav	e_bbae9n.avi	
nine_bbae9n.wav	nine_bbae9n.avi	
now_bbae9n.wav	now_bbae9n.avi	

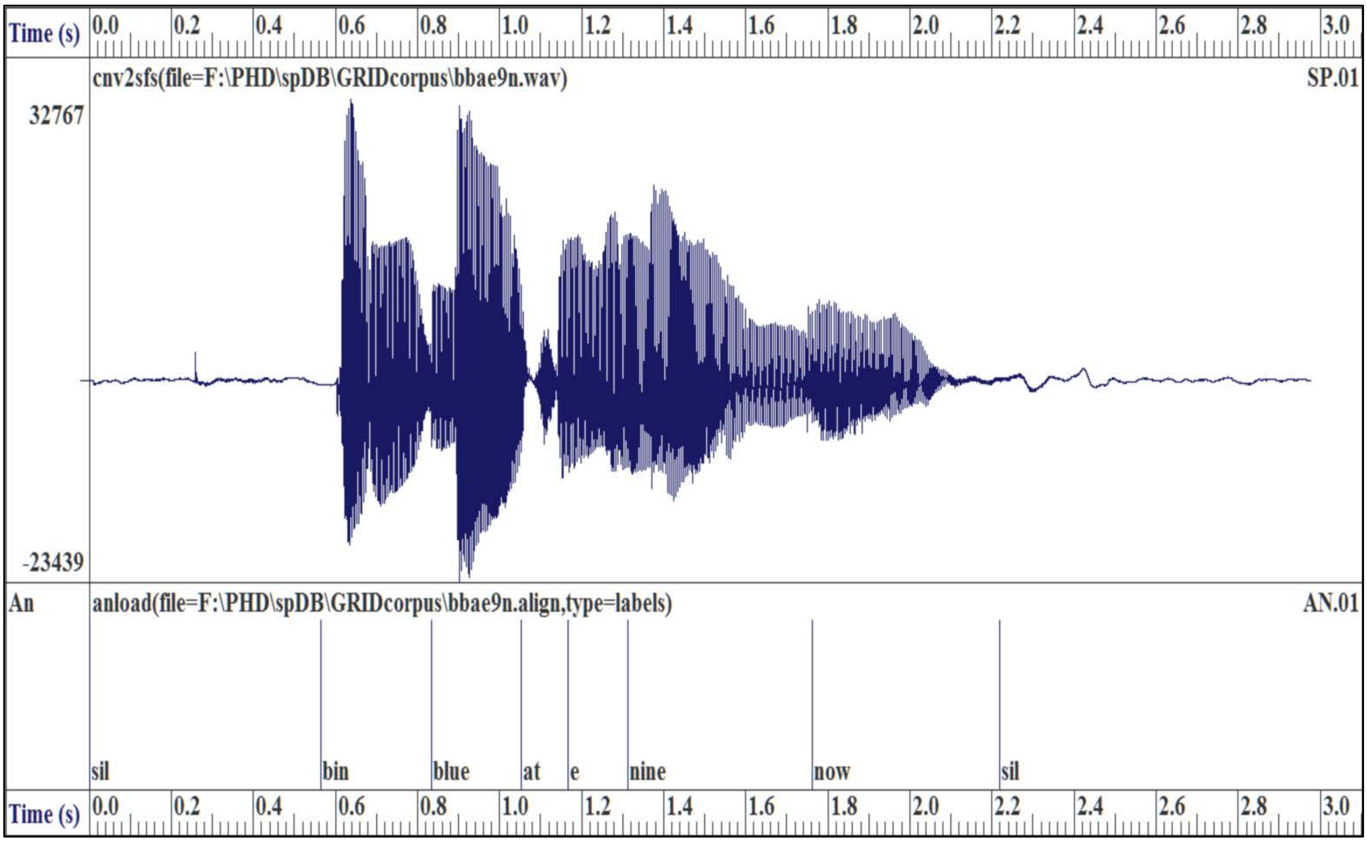

Figure 2: Word boundary according to alignment file, output from SFS [45] program.

B. *Visual Front-End*

This subsection explains the pre-processing steps performed on the captured images from the input video and visual feature extraction operation. These steps are essential in order to obtain an accurate ROI and extract the most effective features.

1) *Visual Pre-Processing:* The pre-processing steps perform framing to divide the isolated word video file to separated images which has a lot of background information which is not useful in the speech recognition task. We use the face-detector module in OpenCV [46] to detect and extract face and ROI for visual features is the speaker's mouth region by using Viola-Jones algorithm [47] from the images. The detected mouth image is converted to grayscale then resized to be (32*32 pixels) in order to make the calculation of DCT features not affected by the lip location in the input image. Figure 3 explains the pre-processing steps for image for speaker 12 from GRID dataset.

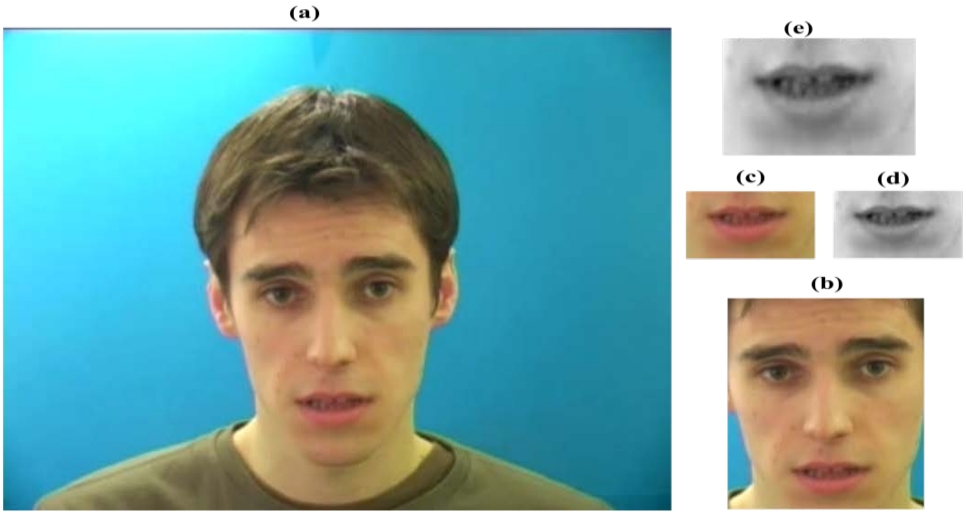

Figure 3: Visual pre-processing steps. a) Original image, b) Detected face, c) Detected mouth, d) Mouth in gray scale, and e) Mouth after resizing 32x32

2) *Visual Features Extraction:* There are two main visual feature extraction classes, appearance or pixel-based and shape or model-based. Model-based feature depend on a geometry dimension of the ROI the width and the height of the speaker's lips, it doesn’t depend on the whole lip region [12] and also doesn’t give the all required information. Appearance-based depends on all pixels in the mouth region and have valuable information to speech recognition [48]. Examples of appearance-based features methods are DCT, Discrete Wavelet Transform (DWT), and Linear Discriminate Analysis (LDA). Due to the good performance of DCT in a lot of previously discussed AVSR systems [49], it is applied in this study. Two-dimensional DCT is applied to the mouth region of the speaker and gives a matrix of features that have the exact dimension as the input mouth image file. Then, extract the final visual feature vector of 13 features from the upper left corner by using zigzag scanning as shown in Figure 4. We try to increase or decrease the visual feature vector around this value but it gives the highest recognition rate.

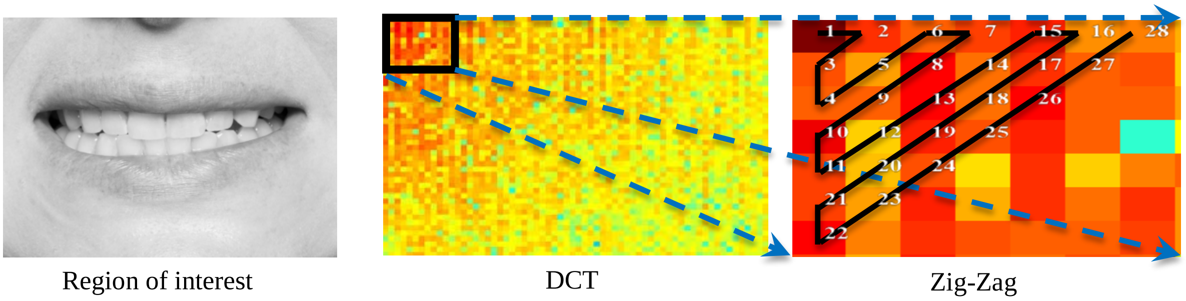

Figure 4: Zigzag scanning to extract feature vector from low frequency components of DCT matrix.

C. Audio Front-End

Before extracting features from the audio file it must be converted to frames, it is popular for speech signal to use frame length not more than 25ms where the speech signal holds its properties [13], then perform frame overlapping to ensure the continuity of the speech signal properties with the adjacent frames, finally frame scaling is done by cross-multiplying the signal by Hamming window. Now data is ready for the feature extraction step, Mel frequency cepstral coefficient (MFCC) is the most effective audio features extraction method [50] which simulate the variation of the human ear’s important bandwidth with frequency. HMM Toolkit (HTK) [51] is used for extracting 13 MFCC features in addition to their 1st and 2nd derivatives which resulting in an acoustic feature vector of length 39 elements.

D. Audio-Visual Features Integration

Integration of audio-visual features can be divided into two categories: early integration (feature fusion), and late integration (decision fusion) as shown in Figure 5. In feature fusion, features from audio and visual are concatenated to form combined single feature vector, which passed after that to the classifier to perform the recognition process. In decision fusion, separated classifier is used for audio and visual part and the output from the different classifier are combined to take the final decision. Here we use early integration but firstly, we need to guarantee identical feature frame rates for audio and visual; the visual features are linearly interpolated to up-sample the video frame rate to have the same frame rate as audio features. Then, the audio and visual features are concatenated and the combined feature vector of dimensionality either 26 (13 from video vector+13 from audio vector) or 52(13 from video vector+39 from audio vector) is used for training and testing the classification stage.

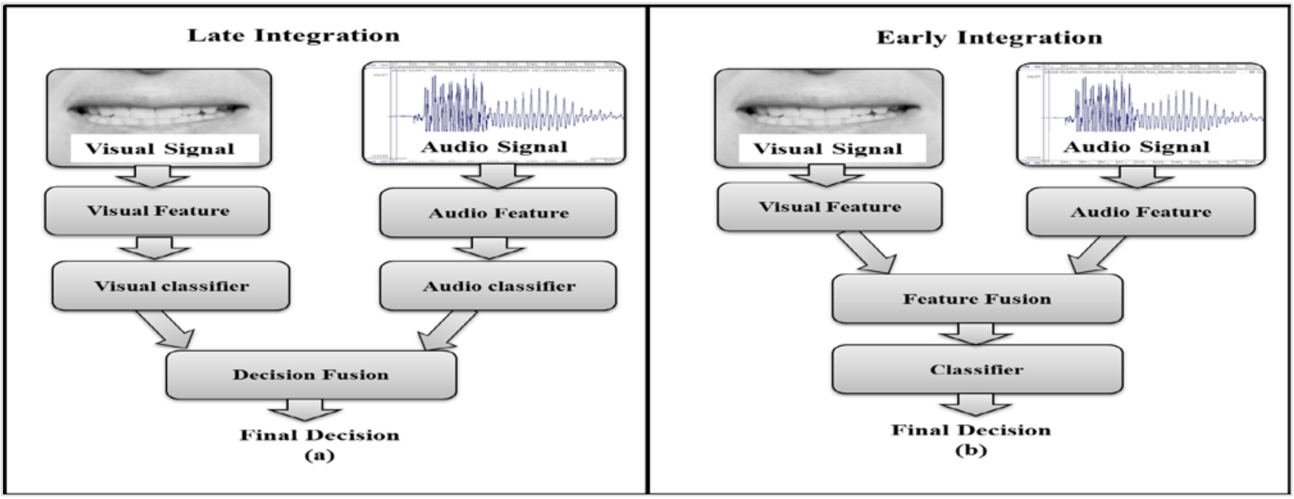

Figure 5: Different techniques for audio- visual integration (a) late integration (b) early integration

E. Classification

Previously HMM was the state of art classification technique for speech recognition system for normal [13] and disorder people [14]. Although, it is easier to understand and implement, now deep learning gives much more accurate results. Deep learning is preferred over HMM due to its robust self-learning mechanism and confirmed performance for speech recognition applications [15]. Deep learning would probably take more computational time, but the results can be more reliable [16]. The accuracy in certain image recognition and language processing problems is superior for deep learning.

In the classification process for frame of speech data, it is much more helpful to look at the frames after it in addition to the previous frames, especially when it occurs close to the end of a word [34] to get the recognized word accurately. RNN is the suitable for these cases, Feng. W [30] explained the main structure of RNN, bidirectional RNNs, and LSTM which we use in building the classifier for our proposed AVSR model. Although RNN have some problems: firstly, since they treat inputs in temporal order, their outputs generally based on previous context; secondly they have trouble learning time-dependencies more than a few timesteps long as mentioned in [34] ,in addition to, it facing a challenging problem known as the gradient vanishing and exploding problem [52]. The solution for these problems is to use bidirectional LSTM [53] [54]. LSTM model firstly proposed by Hochreiter and Schmidhuber [55], it confirms that the gradients can pass through many time steps and overcomes both vanishing and exploding insufficiencies of the gradients [30]. Bidirectional recurrent neural (BiRNN) networks introduced by Schuster et al. [56] , it consists of two RNNs forward and backward ones. Because of its good results, BiRNN has been used in speech recognition [54], [57], and handwriting recognition [58], [59] systems. In combination with the benefits of BiRNN and the enhancement that LSTM introduces [54] BiLSTM can be the optimal solution for the classification process in speech recognition. The structure of BiLSTM is shown in Figure 6. Because of its great enhancement proved by the previously system, we decide to use BiLSTM in compared to HMM for the classification process in our model.

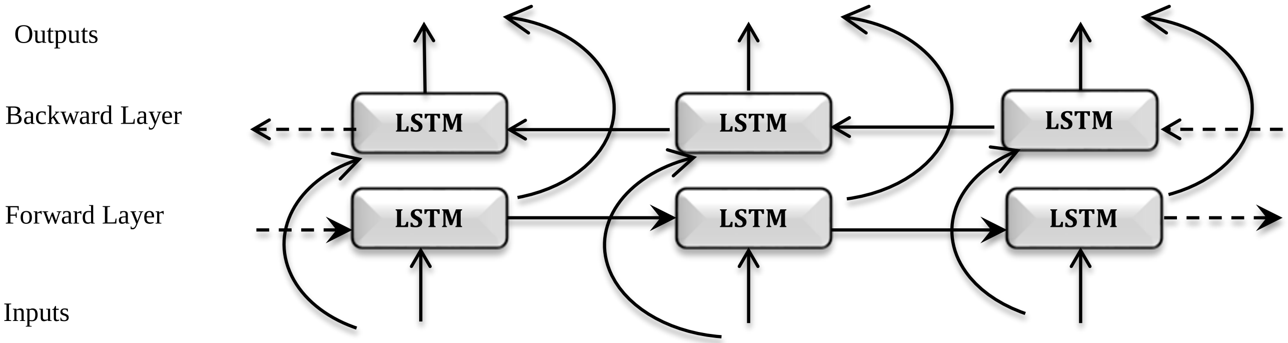

Figure 6: Bidirectional Long Short-Term Memory Network (BiLSTM)

In this paper, we compare the obtained result from BiLSTM classifier with HMM to get the optimal accuracy for our recognition system. In case of HMM classifier HMM toolkit (HTK) [51] is used for organizing, training and testing the HMM model. A total of 51 HMM models, one for each word, are trained for Grid database. The proposed model uses 5-state with various numbers of Gaussian mixtures from 2 to 128 mixtures. In order to select the optimal number of mixtures, it is a good idea to gradually increase the number of mixtures by two. The step by step increasing allows recognition performance to be monitored to find the optimal number of mixtures which gives the best recognition accuracy results.

4 PROCEDURE AND SYSTEM MODEL

This section explains the dataset used in training and validating the proposed model, and the performance analysis techniques that used to evaluate the obtained recognition results

A. Dataset

GRID audio-visual corpus is used in the training and the testing stages. It is a collection of audio and video recordings of 34 speakers (18 male, 16 female) ages ranged from18 to 49 years each saying 1000 sentences [32]. The total length of the recordings is 28 hours; with total number of words in the vocabulary are 51. The syntactic structures of all sentences are the same as shown below.

<command>< color >< preposition >< letter >< digit >< adverb > [13]

The vocabulary of the GRID corpus consists of 4 command words, 4 words representing color, 4 prepositions, 26 letters, 10 digits, and 4 adverbs as listed in Table II. The video was recorded as a sequence of images with a frame rate of 40ms. Figure 7 shows the grammar file for the GRID corpus it has a similar construction as given in [13], but here for isolated word not for a complete sentence, figure 8 gives example frame for GRID dataset.

TABLE II

SENTENCE STRUCTURE FOR THE GRID CORPUS [32]

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

```
$COMMAND= BIN | LAY | PLACE | SET;
$COLOR= BLUE | GREEN | RED | WHITE;
$PREPOSITION= AT | BY | IN | WITH;
$LETTER=A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | X | Y | Z;
$DIGIT = ZERO | ONE | TWO | THREE | FOUR | FIVE | SIX | SEVEN | EIGHT | NINE;
$ADVERB= AGAIN | NOW | PLEASE | SOON;
( SENT-START ($COMMAND | $COLOR | $PREPOSITION | $LETTER | $DIGIT | $ADVERB) SENT-END)
```

Figure 7: Grammar file [13] for GRID corpus

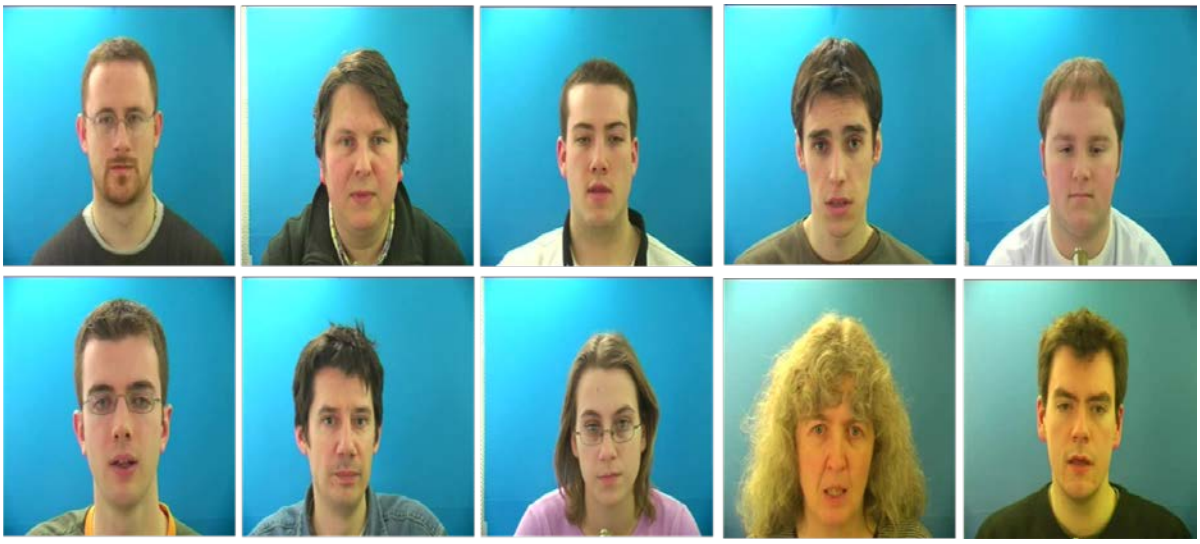

Figure 8: Example frame from different speakers in the GRID dataset

B. Performance analysis

The system performance is analyzed, in case of HMM classifier, by using Hresults from HTK tool to calculate the recognition accuracy of the speech system, which evaluated as % Accuracy=100 x (N-D-S-I)/N [51] where N total number of words in test set, D number of deletions, S number of substitutions, I number of insertions and H number of correct labels.

5 Results

This section introduces an evaluation of the proposed model and compares our methods to the previous state of the art. The results for GRID dataset will be introduced here for audio only with different feature vector sizes, visual only, and early integrated audio-visual features either in clean media or after adding babble noise with 5db signal-to-noise ratio.

C. GRID Results:

This subsection presents the results of using the GRID dataset in testing the proposed model. As mentioned before, each video in the Grid dataset represents a sentence with 6 words, in order to perform isolated word recognition we segmented the audio and video file to word boundary. Therefore, we select four speakers from GRID to test our system, each speaker will have 6000 video files, and for the four speakers we have 6000*4=24000 videos. For each speaker we take 75% for training and 25% for testing, audio features are extracted by using MFCC with feature vector of size 13 or 39, and DCT to extract the visual features with feature vector of size 13, audio-visual features obtained by concatenating both feature vectors.

To precisely compare our results with [31], we initially performed our experiments on speaker four (S4, female) as done there; we also perform experiments on speakers 6, 11, and 12 (two females and two males). Table IV lists the recognition accuracies of using BiLSTM, and HMM with different Gaussian mixtures. It also shows the results for clean data and after adding babble noise with SNR 5db. The grey cells refer to the enhancement in recognition accuracy that occurs after adding visual features. Figure 9, 10, 11, and 12 compares the results of the three classifiers to identify the best feature type and best classifier. Figure 13 introduces the confusion matrix for speaker 12, using BiLSTM for AVSR system after adding the visual feature of size 13 DCT to audio MFCC with size 39. The reference labels are represented in rows, and classification postulate is represented in columns. The recognition accuracy for longer words like "please" is greater than for single letter like "A".

From the obtained result illustrated in table IV we found that:

- Using HMM with integrated audio-visual feature enhanced the recognition accuracy over audio-only by 10.58%, 3.35%, 2.27%, and 1.45% in clean environment, while after adding babble noise enhancement is 12.05%, 3.35%, 2.27%, and 1.45% for speaker 4, 6 ,11 and 12 respectively.
- When using BiLSTM, the best accuracy occurred when using integrated audio-visual feature in clean and noisy environment for speaker 4,6,11 and 12.

Our proposed model gives recognition accuracy for speaker 4 is 99.07% with BiLSTM classifier which is better than 79.9% obtained by Ephrat [31] for the same dataset and same speaker, the comparison between our model result and Ephrat [31] is illustrated in table III. It also gives better results than using HMM for the other tested speakers.

TABLE III

COMPARISON BETWEEN THE BEST RECOGNITION ACCURACY OBTAINED by OUR MODEL and THE PREVIOUSLY OBTAINED RESULTS in [31]

	Ephrat [31]	**Ours**
Audio-only	82:6%	98.87%
Audio-visual	79.9%	99.07%

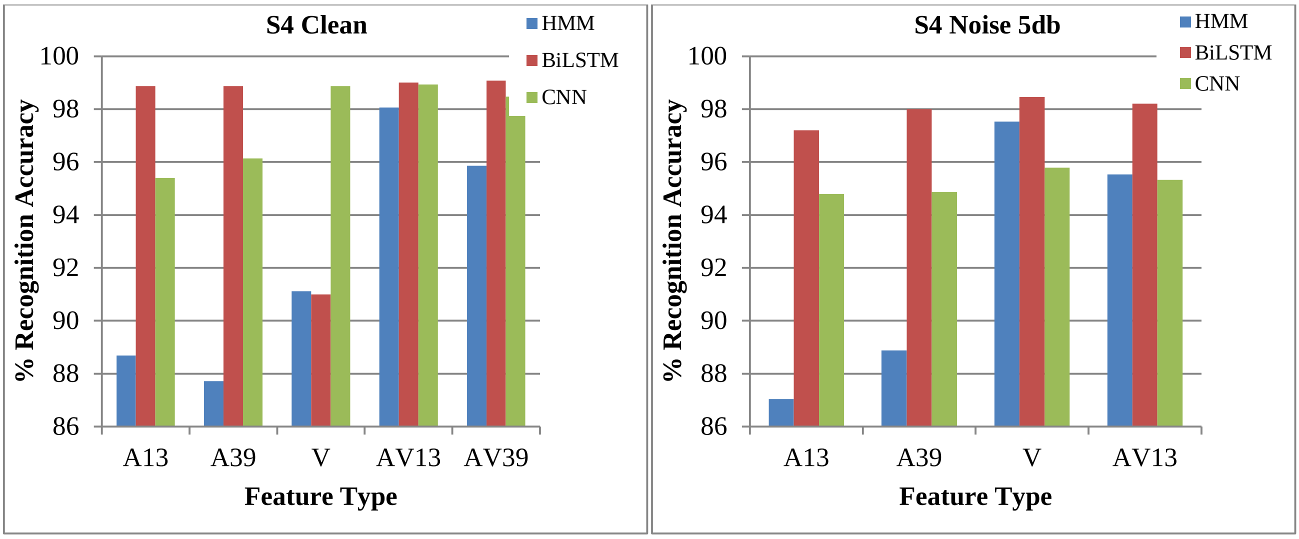

Figure 9: %Recognition accuracy results for speaker 4 with different classifiers in clean and noisy environment

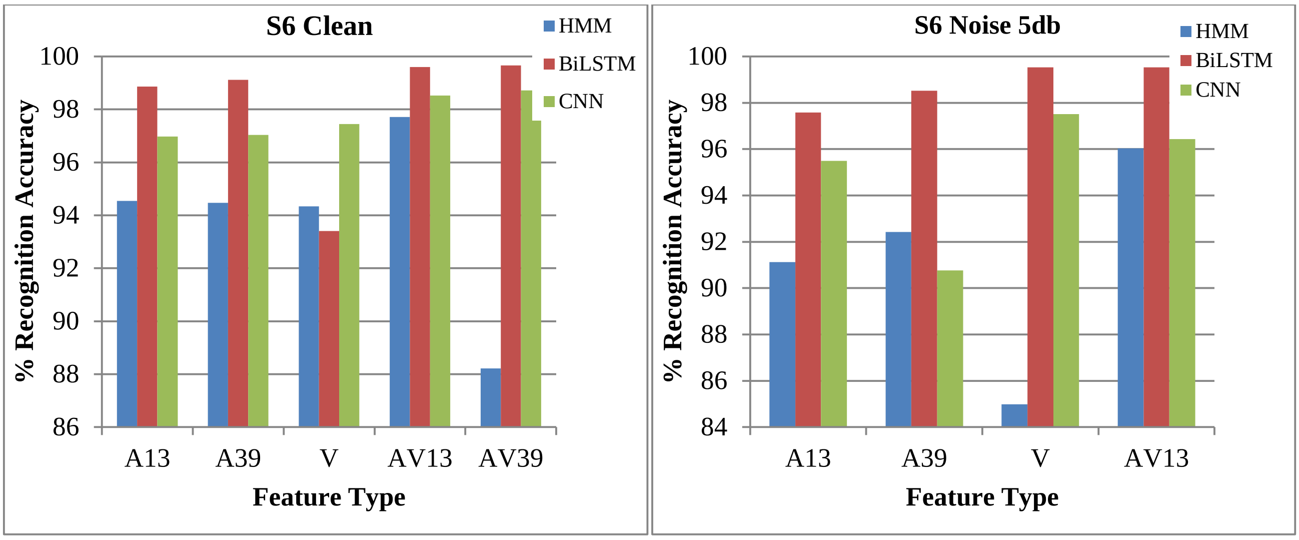

Figure 10: %Recognition accuracy results for speaker 6 with different classifiers in clean and noisy environment

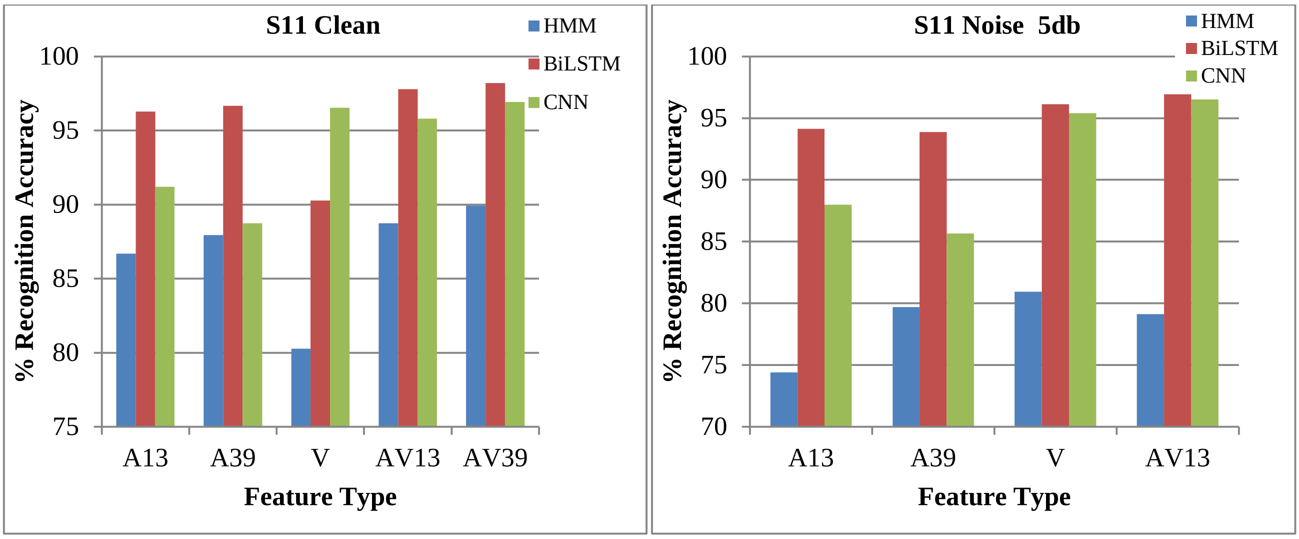

Figure 11: % Recognition accuracy results for speaker 11 with different classifiers in clean and noisy environment

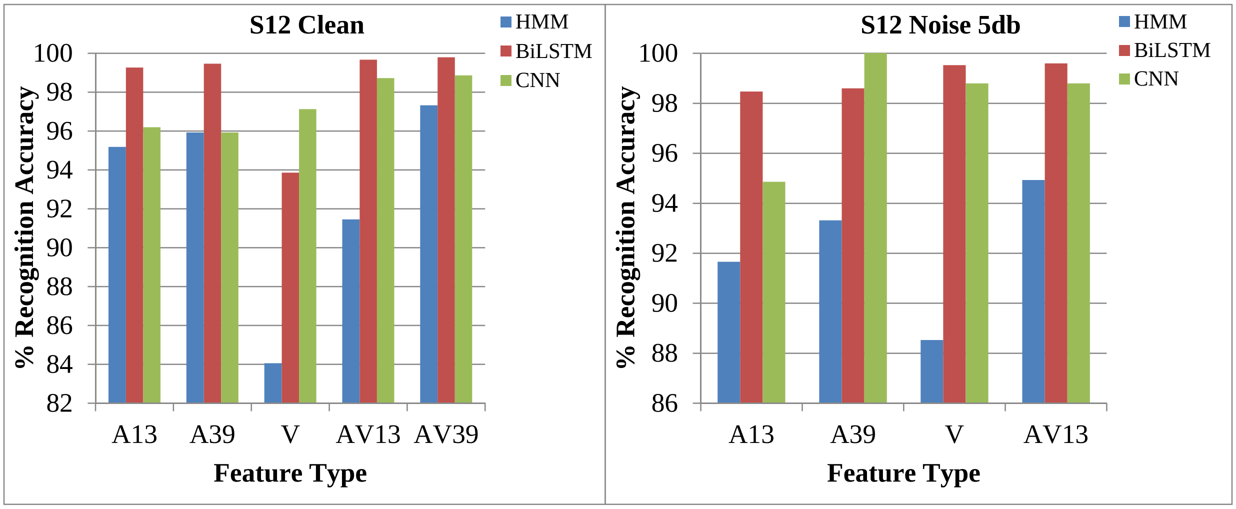

Figure 12: %Recognition accuracy results for speaker 12 with different classifiers in clean and noisy environment

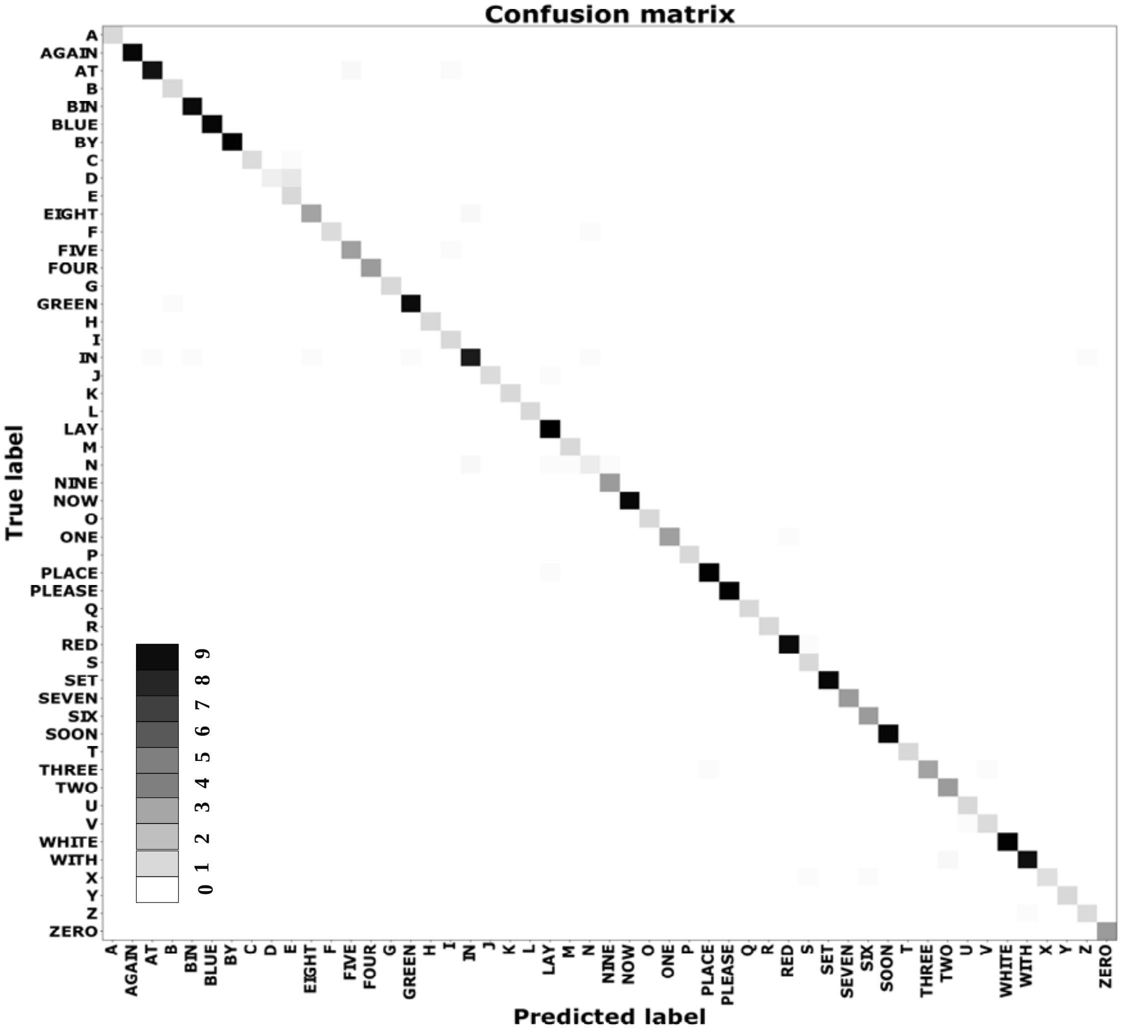

Figure 13: GRID confusion matrix for speaker 12, with audio-visual feature vector size of 52

Figure 13: GRID confusion matrix for speaker 12, with audio-visual feature vector size of 52

TABLE IV
ACCURACY RESULTS for SPEAKERS 4, 6, 11, and 12 for GRID DATASET

	SNR		Accuracy			
			S4	S6	S11	S12
BiLSTM	Clean	A13	97.75%	98.86%	96.27%	99.27%
		A39	98.87%	99.12%	96.67%	99.47%
		V13	90.99%	93.40%	90.27%	93.86%
		AV13	99.00%	99.60%	97.80%	99.67%
		AV39	99.07%	99.66%	98.20%	99.80%
	5db	A13N	97.20%	97.58%	94.13%	98.47%
		A39N	98.00%	98.52%	93.87%	98.60%
		AV13N	98.46%	99.53%	96.13%	99.53%
		AV39N	98.20%	99.53%	96.93%	99.60%
HMM	Clean	A13	88.68 mix16	94.54 mix16	86.69 mix8	95.19 mix8
		A39	87.72 mix16	94.47 mix16	87.93 mix8	95.93 mix8
		V13	91.12 mix16	94.34 mix16	80.27 mix8	84.06 mix8
		AV13	98.06 mix16	97.71 mix16	88.73 mix8	91.46 mix8
		AV39	95.86 mix16	88.22 mix16	89.93 mix8	97.33 mix8
	5db	A13N	87.04 mix8	91.12 mix16	74.42 mix8	91.66 mix4
		A39N	88.88 mix8	92.42 mix16	79.7 mix8	93.32 mix4
		AV13N	97.53 mix8	84.98 mix16	80.93 mix8	88.53 mix4
		AV39N	95.53 mix8	96.03 mix16	79.13 mix8	94.93 mix4

6 CONCLUSION

In this work, Audio-visual model (BiLSTM-AVSR) based on deep learning approach is presented as a solution for speech recognition system. The proposed model implies segmenting the input video to isolated word boundary, then extracting features from audio and video files using MFCC and DCT separately. After that, early integration is used to concatenate the obtained audio and visual features in a single feature vector. Finally, BiLSTM is used in the classification process in comparison with traditional HMM. The model is evaluated using multi-speakers GRID audio-visual dataset with dependent speaker experiments. Based on the obtained results, we can conclude that increasing the size of audio feature vector from 13 to 39 doesn't give effective enhancement for the recognition accuracy in clean environment, but in noisy environment, it gives better performance. Combining visual feature to audio feature increase the recognition accuracy. BiLSTM is considered to be the optimal classifier for a robust speech recognition system when compared to traditional HMM, because it takes into consideration the sequential characteristic of the speech signal (audio and visual). The proposed model gives an increment in the recognition accuracy and decreasing in the loss value for both clean and noisy environments overusing audio-only. Comparing the proposed model to previously obtain results using the same dataset, we found that our model gives higher recognition accuracy.

REFERENCES

[1] F. Tao and C. Busso, "Lipreading Approach for Isolated Digits Recognition Under Whisper and Neutral Speech," in *Proc. Interspeech*, 2014, pp. 1154 – 1158.

[2] T. F. Cootes, J. A. Bangham, S. Cox, R. Harvey I. Matthews, "Extraction of visual features for lipreading," *IEEE Transactions onPattern Analysis and Machine Intelligence*, pp. 198-213, 2002.

[3] M. Barnard, M. Pietikainen G. Zhao, "Lipreading with local spatiotemporal descriptors," *IEEE Transactions on Multimedia*, vol. 11, pp. 1254–1265, 2009.

[4] E. D. Petajan, "Automatic Lipreading to Enhance Speech Recognition (Speech Reading)," *Ph.D. dissertation, University of Illinois at Urbana-Champaign*, 1984.

[5] C., Potamianos, G., Luettin, J., Matthews, I., Glotin, H., Vergyri, D.,& Zhou, J. Neti, "Audio-visual speech recognition final workshop," *Center for Language and Speech Processing, Johns Hopkins University, Baltimore.*, 2000.

[6] G., Graf, H. P., & Cosatto, E. Potamianos, "An image transform approach for HMM based automatic lipreading.," in *In Proceedings International Conference on Image Processing. ICIP98 (Cat. No. 98CB36269). IEEE.*, 1998, pp. 173-177.

[7] G., Neti, C., Iyengar, G., Senior, A. W., & Verma, A. Potamianos, "A cascade visual front end for speaker independent automatic speechreading," *International Journal of Speech Technology, 4(3-4),* pp. 193-208, 2001.

[8] Y. Yamaguchi, K. Nakadai, H. G. Okuno, T. Ogata. K. Noda, "Lipreading using convolutional neural network," *INTERSPEECH*, pp. 1149–1153, 2014.

[9] C. L. Chowdhary, "Application of Object Recognition With Shape-Index Identification and 2D Scale Invariant Feature Transform for Key-Point Detection," *In Feature Dimension Reduction for Content-Based Image Identification. IGI Global.*, pp. 218-231, 2018.

[10] M. T. Chan, "HMM-based audio-visual speech recognition integrating geometric-and appearance-based visual features.," *IEEE Fourth Workshop on Multimedia Signal Processing (Cat. No. 01TH8564) IEEE.*, pp. 9-14, October 2001.

[11] H. McGurk , J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.

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

[13] Amr M. Gody , M. Hesham Farouk Eslam E. El Maghraby, "Enhancing quality and accuracy of speech recognition system by using multimodal audio-visual speech signal," *12th International Computer Engineering Conference (ICENCO)*, pp. 219-229, 2016.

[14] Reda A. El-Khoribi, Mahmoud E. Shoman Elham S. Salama, "Audio-Visual Speech Recognition for People with Speech Disorders," *International Journal of Computer Applications (0975 –8887)*, vol. 96–No.2, June 2014.

[15] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, *61*, pp. 85-117, 2015.

[16] Geoffrey, Li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior et al. Hinton, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal processing magazine 29*, 2012.

[17] & Acharjya, D. P. Chowdhary C. L., "Singular Value Decomposition–Principal Component Analysis-Based Object Recognition Approach," *Bio-Inspired Computing for Image and Video Processing*, pp. 323-341, 2018.

[18] C. L. Chowdhary, "Linear feature extraction techniques for object recognition: study of PCA and ICA," *Journal of the Serbian Society for Computational Mechanics, 5*, pp. 19-26, 2011.

[19] C. L. Chowdhary, "3D object recognition system based on local shape descriptors and depth data analysis," *Recent Patents on Computer Science, 12*, pp. 18-24, 2019.

[20] S. Petridis , M. Pantic., "Deep complementary bottleneck features for visual speech recognition," *ICASSP*, pp. 2304–2308, 2016.

[21] Wei Li, Yifan Zhang, and Zhiyong Feng Fan Zhang, "Data Driven Feature Selection for Machine Learning Algorithms in Computer Vision," *IEEE Internet of Things Journal*, pp. 4262-4272, 2018.

[22] H. Ney, R. Bowden O. Koller, "Deep learning of mouth shapes for sign language," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 85–91, 2015.

[23] Alan J., Oscar N. Garcia, Eric D. Petajan Goldschen, "Continuous automatic speech recognition by lipreading," *Springer,Dordrecht*, pp. 321-343, 1997.

[24] H. Ninomiya, N. Kitaoka, S. Osuga, Y. Iribe,K. Takeda, S. Hayamizu S. Tamura, "Audio-visual speech recognition using deep bottleneck features and high performance lipreading.," *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (AP-SIPA)*, pp. 575–582, 2015.

[25] F. Makedon. G. Galatas, "Audio-visual speech recognition incorporating facial depth information captured by the kinect.," *Signal Processing Conference(EUSIPCO), Proceedings of the 20th European*, pp. 2714–2717, 2012.

[26] K., Yamaguchi, Y., Nakadai, K., Okuno, H. G., & Ogata, T. Noda, "Audio-visual speech recognition using deep learning," *Applied Intelligence 42.4*, pp. 722-737, 2015.

[27] E. Marcheret, V. Goel. Y. Mroueh, "Deep multimodal learning for audio-visual speech recognition," *IEEEInternational Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 2130–2134, 2015.

[28] T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, M. Pantic S. Petridis, "End-to-end audiovisual speech recognition," *CoRR, abs/1802.06424*, 2018.

[29] T. Stafylakis , G. Tzimiropoulos., "Combining residualnetworks with LSTMs for lipreading," *Interspeech*, 2017.

[30] Guan N, Li Y, Zhang X, Luo Z. Feng W, "Audio visual speech recognition with multimodal recurrent neural networks," *2017 International Joint Conference on Neural Networks (IJCNN). IEEE*, pp. 681-688, 2017.

[31] A., Peleg, S. Ephrat, "Vid2speech: Speech reconstruction from silent video," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*, 2017.

[32] J. Barker, S. Cunningham, X. Shao M. Cooke, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Societyof America, 120(5)*, pp. 2421–2424, 2006.

[33] P. E., Mun, H. K., & Vaithilingam, C. A. James, "A Hybrid Spoken Language Processing System for Smart Device Troubleshooting," *Electronics, 8(6), 681.*, 2019.

[34] A., Fernández, S., & Schmidhuber, J. Graves, "Bidirectional LSTM networks for improved phoneme classification and recognition," in *International Conference on Artificial Neural Networks. Springer*, Berlin, 2005, pp. 799-804.

[35] M., Koutník, J., Schmidhuber, J. Wand, "Lipreading with long short-term memory," *ICASSP'16*, pp. 6115–6119, 2016.

[36] J. S., Senior, A., Vinyals, O., Zisserman, A Chung, "Lip reading sentences in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3444-3453.

[37] Thanda, Shankar M. Venkatesan Abhinav, "Audio Visual Speech Recognition using Deep Recurrent Neural Networks," *IAPR Workshop on Multimodal Pattarn Recognition of Social Signals in Human-Computer Interaction, Springer,Cham*, pp. 98-109, 2016.

[38] B. Shillingford, S. Whiteson, N. de Freitas Y. M. Assael, "Lipnet: Sentence-level lipreading," in *GPU Technology Conference*, 2016.

[39] S. Fern´andez, F. Gomez, J. Schmidhuber. A. Graves, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," *Proceedings of the International Conference on MachineLearning*, pp. 369–376, 2006.

[40] J., Vincent, E., Ma, N., Christensen, H., & Green, P Barker, "The PASCAL CHiME speech separation and recognition challenge," *Computer Speech & Language, 27(3)*, pp. 621-633, 2013.

[41] T., Menzel, W., & Yang, S Gan, "An audio-visual speech recognition framework based on articulatory features," 2007.

[42] B., & Le Cornu, T. Milner, "Reconstructing intelligible audio speech from visual speech features," *Interspeech* , 2015.

[43] Richard Harvey Helen L Bear, "Decoding visemes:Improving machine lip-reading," *ICASSP'16*, pp. 2009-2013, 2016.

[44] MATLAB program. [Online]. http://www.mathwork.com

[45] Huckvale Mark. (2008) Speech filing system. [Online]. www.phon.ucl.ac.uk/resources/sfs

[46] Intel Corporation Willow Garage. OpenCV. [Online]. www.opencv.org

[47] O. H. Jensen, "Implementing the Viola-Jones face detection algorithm," (Master's thesis, Technical University of Denmark, DTU, DK-2800 Kgs. Lyngby, Denmark), 2008.

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

[49] V., Thiran, J. P. Estellers, "Multi-pose lipreading and audio-visual speech recognition," *EURASIP Journal on Advances in Signal Processing*, 2012.

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

[51] S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Valtchev, V. Young, *The HTK Book*, 341st ed.: Cambridge University Press., 2006.

[52] P. Simard, P. Frasconi Y. Bengio, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994.

[53] M. Schuster and K. K. Paliwal., "Bidirectional recurrent neural networks. ," *IEEE Transactions on Signal Processing, 45*, pp. 2673–2681, November 1997.

[54] J. Schmidhuber A. Graves, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *NeuralNetworks*, vol. 18, no. 5, pp. 602–610, 2005.

[55] S. Hochreiter , J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[56] M. Schuster , K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[57] A. r. Mohamed, G. Hinton A. Graves, "Speech recognition with deep recurrent neural networks," *2013 IEEE international conference on acoustics, speech and signal processing.IEEE*, pp. 6645–6649, 2013.

[58] M. Liwicki, H. Bunke, J. Schmidhuber, and S. Ferńandez A. Graves, "Unconstrained on-line handwriting recognition with recurrent neural networks," *Advances in Neural Information Processing Systems*, pp. 577–584, 2008.

[59] M. Liwicki, S. Ferńandez, R. Bertolami, H. Bunke, J. Schmidhuber A. Graves, "A novel connectionist system for unconstrained hand-writing recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 5, pp. 855–868, 2009.

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 M.Sc 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 B.sc (Honours) degree in communication and electronics from faculty of engineering, Fayoum University in 2008. She received the M.sc 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

التعرف على الكلام باستخدام النهج التاريخي المتعدد الوسائط

إسلام عيد علي محمد المغربي* ، عمرو محمد رفعت جودي*، محمد هشام فاروق**

*قسم هندسة الإتصالات والإليكترونيات - كليه الهندسة - جامعة الفيوم

**قسم هندسة الرياضيات والفيزيقا - كليه الهندسة - جامعة القاهره

[1]eem00@fayoum.edu.eg

[2]amg00@fayoum.edu.eg

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

ملخص:

يقوم هذا البحث بتصميم نظام للتعرف علي الاصوات معتمداً علي الاشارة الصوتية بالاضافة الي الاشارة البصريه المصاحبة للصوت المأخوذه من حركة الشفاه للمتكلم. الدراسات السابقة اثبتت ان إضافة حركة الشفاه إلي الإشاره الصوتيه من الممكن أن يؤدي إلي زيادة دقة التعرف علي الاصوات خاصة في حالة وجود ضوضاء مؤثره علي الإشاره الصوتيه. . من خلال هذا النظام المقترح في هذا البحث يتم إستخراج خصائص الإشارة الصوتيه بإستخدام خاصية MFCC وإستخدام خاصية DCT لإستخراج الخصائص من صورة حركة الشفاه المصاحبة للصوت. . يتم دمج الخصائص المستخرجه من الصوت والصوره المصاحبه له لتدريب نظام التعرف علي الاصوات باستخدام واحدة من اهم انواع Deep Learning ألا وهي BiLSTM التي تتميز بكفاءتها في تصنيف الإشاره الصوتيه لأنها تأخذ في إعتبارها جميع خصائص الصوت . هذا البحث يجري مقارنه بين النتائج التي تم التوصل إليها من خلال إستخدام BiLSTM للتعرف علي الصوت بإستخدام الخصائص التي تم إستخراجها من الصوت والصوره معاً والنتائج التي تم الحصول عليها بإستخدام الطريقة المستخدمه من قبل لتصنيف الاصوات وهي HMM عن طريق استخدام اداة HTK. تم اختبار كفاءة النظام المقترح من خلال تطبيقة باستخدام واحدة من اكبر قواعد البيانات للصوت والصورة معا وهي قاعدة بيانات GRID. من خلال تحليل النتائج للنظام المقترح المعتمد علي الصوت والصوره معا وإستخدام BiLSTM في مرحلة التصنيف نجد انه يقدم كفاءة اعلي ومعدل تعرف اكبر بمقدار 9.28% عن استخدام الصوت فقط.

الكلمات المفتاحية: ***DCT, MFCC, HMM, BiLSTM, and GRID.***