Noise-Robust Speech Recognition System based on Multimodal Audio-Visual Approach Using Different Deep Learning Classification Techniques

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

*Electrical Engineering Department, Faculty of Engineering, Fayoum University,
Fayoum, Egypt
[1]eem00@fayoum.edu.eg
[2]amg00@fayoum.edu.eg

**Engineering Math. & Physics Dept., Faculty of Engineering, Cairo University,
Cairo, Egypt
[3]mhesham@eng.cu.edu.eg

Abstract: ***Multimodal speech recognition is proved to be one of the most promising solutions for designing a robust speech recognition system, especially when the audio signal is affected by noise. The visual signal can be used to obtain more information to enhance the recognition accuracy in a noisy system, whereas the reliability of the visual signal is not affected by the acoustic noise. The critical stage in designing a robust speech recognition system is the choice of an appropriate feature extraction method for both audio and visual signals and the choice of a reliable classification method from a large variety of existing classification techniques. This paper extends an earlier work on designing a speech recognition system based on Hidden Markov Model (HMM) classification technique of using visual modality in addition to audio modality[1]. Improved off traditional HMM-based Automatic Speech Recognition (ASR) accuracy is achieved by implementing a technique using either RNN-based or CNN-based approach. This research is intending to deliver two contributions: The first contribution is the methodology of choosing the visual features by comparing different visual features extraction methods like Discrete Cosine Transform (DCT), blocked DCT, and Histograms of Oriented Gradients with Local Binary Patterns (HOG+LBP), and applying different dimension reduction techniques like Principal Component Analysis (PCA), auto-encoder, Linear Discriminant Analysis (LDA), t-distributed Stochastic Neighbor Embedding (t-SNE) to find the most effective features vector size. Then the obtained visual features are early integrated with the audio features obtained by using Mel Frequency Cepstral Coefficients (MFCCs) and the combined audio-visual feature vector is fed to the classification process. The second contribution of this research is the methodology of developing the classification process using deep learning by comparing different Deep Neural Network (DNN) architectures like Bidirectional Long-Short Term Memory (BiLSTM) and Convolution Neural Network (CNN) with the traditional HMM. The proposed model is evaluated on two multi-speakers AV-ASR datasets named AVletters[1] and GRID[2] with different SNRs. The model performs speaker-independent experiments in AVlettter dataset and speaker-dependent in GRID dataset. The experimental results obtained in this research showed that using early integration between audio features obtained by MFCC and visual features obtained by DCT with zigzag scanning demonstrate higher recognition accuracy when used with BiLSTM classifier compared to other methods for features extraction, dimension reduction, and classification techniques.***

Keywords: ***AV-ASR, DCT, Blocked DCT, PCA, MFCC, HMM, DNN, AVletters, and GRID.***

1 INTRODUCTION

The main goal of designing a speech recognition system is to obtain high quality and robust model, especially in a noisy environment. Speech is a multimodal signal that depends on audio and visual modalities, so to build high quality and noise-robust speech recognition system it is important to take advantage of the different modalities of the speech signal to enhance the speech understanding process. Using visual modality like lips movements to identify the spoken words called lipreading can add additional information about the recognized word in clean and noisy environments [2], [3] because it is not affected by the acoustic noise. Lipreading is considered as a significant research topic that attracts researchers attention for the last decades, especially when used in some applications such as sign language recognition [4], hearing aids [5], speaker recognition [6], and also can be used to enhance speech recognition performance [7] when used with the audio signal. Combining the achievement of the lipreading to the traditional audio-only automatic speech recognition system is a good choice for designing a noise-robust AV-ASR system. AV-ASR can be used for noisy system and overlapped speech recognition which considered a highly challenging task to date [8]. The main parts of the building AV-ASR system are explained in [9].

The first step in building noise-robust speech recognition is to carefully choose the suitable method for extracting the most informative feature from the audio and visual signals. Feature extraction is considered to be an important issue for designing the recognition system. Extracting the features from the visual signal can be divided into geometric feature-

[1] Available online: http://www2.cmp.uea.ac.uk/~bjt/avletters/

[2] Available online: http://spandh.dcs.shef.ac.uk/gridcorpus/

based approaches, Appearance-based approaches, Image-transformed-based approaches, and Hybrid approaches [4]. In this research we used image-transformed-based approach which is transforming the ROI which here is the mouth image to a space of features to perform redundant data elimination. This approach used some transform techniques like DCT [2], Discrete Wavelet Transform (DWT), PCA [10], HOGs and LBPs [11] and linear discriminate analysis (LDA) [12]. DCT proves its effectiveness for visual speech recognition [13] in a lot of researches [14][15]. According to audio feature extraction MFCC is considered to be the most widely used in the speech community, as they provide the best encoding of those audio bands which are most related to the spoken word [16]. Therefore our model uses MFCC for audio feature extraction.

The fusion between the information coming from different sources, visual and audio features, plays an important role in the enhancement of the recognition performance. Finding correlations between different modalities, and modelling their interactions, has been addressed in various learning frameworks and has been applied to AV-ASR. In [17] methods are proposed in deep multimodal learning for fusing speech and visual modalities for Audio-Visual Automatic Speech Recognition (AV-ASR). The work introduced in [18] proposes multimodal attention based method for audio-visual speech recognition which automatically learn the fused representation from both modalities based on their importance by using sequence-to-sequence (Seq2seq) architectures.

The second important step in designing a speech recognition system is the classification process. HMM is considered to be the most widely used classifier in the AV-ASR literature[19][20][21]. In Spite of the HMM efficiency and simplicity, the recent researches prove that using deep learning can give better recognition performance when compared to HMM, because of its self-learning technique and reliable results [22]. The past decade has seen rapid developments of deep learning techniques with significant impacts on signal and information processing. In contrast to traditional machine learning and artificial intelligence approaches, the deep learning technologies have recently been progressing massively with successful applications to audio visual speech recognition [23]. Also the work introduced in [24] proposed an audio-visual deep CNNs (AVDCNN) speech enhancement model, which incorporates audio and visual streams into a unified network model. It also proposes a multi-task learning framework for reconstructing audio and visual signals at the output layer, confirming its capability of effectively combining audio and visual information in speech enhancement.

In this paper, different classification methods are used like HMM and comparing it to the result obtained by using deep learning classification like CNN and BiLSTM to choose the classification method which gives higher recognition accuracy.

There are important tasks in designing a noise-robust speech recognition system that has up till now not be successfully addressed: 1) Choosing the informative feature from the visual speech signal, and suitable methods for reducing the visual feature size. 2) Selecting the data fusion methods which combine the extracted feature from the audio and visual signal, and 3) Comparing the performance of different classification methods.

A lot of Machine Learning algorithms can be used in the classification process like KNN, SVM, logistic regression, etc.., These algorithms are learning less when compared to CNN and RNN; that is because there is no transfer learning happening in Machine learning algorithm, while there is transfer learning in deep learning. This enables Deep learning to learn more, and fewer errors will occur.

This paper extends our previous work introduced in [1] by firstly comparing different visual feature extraction DCT or Blocked DCT, secondly studying the effect of using different dimension reduction techniques like PCA, LDA, t-SNE, and auto-encoder in selecting the most effective visual feature and avoid redundancy of data. Finally due to the effect of CNN proved by the recent studies in speech recognition, so in this research comparing different classification methods like CNN, BiLSTM, and HMM. Also this paper introduces speaker dependent and independent experiments in two different multi-speaker datasets AVLetter and GRID to ensure the robustness of the proposed speech recognition system.

The previous researches gave a high accuracy speech recognition system but searching for more improvement to get a reliable speech system still needs more work. Expanding on ideas from the previously mentioned researches and achievements, our decision is:

- Testing the proposed model with multi-speakers datasets like GRID which is suitable for the learning stage of the neural network, using different size datasets to ensure its efficiency, and perform speaker-dependent and independent experiments.
- Choosing the most functionality feature extraction methods for the visual signal from a large variety of methods, followed by selecting the dimension reduction techniques.
- Integrating the visual speech feature with the acoustic features to design a reliable audio-visual speech recognition system.
- Applying one of the most important Deep Neural Network (DNN) architecture BiLSTM in the classification process, and comparing the obtained results with CNN and HMM.

The next section describes the proposed system which uses different classification techniques to test the enhancement that the visual features introduced to the recognition accuracy.

2 PROPOSED MODEL

The pipeline of the proposed AV-ASR model is shown in figure 1 which explains the main stages for the proposed system including Data preparation, Pre-processing, Feature Extraction, Feature Interpolation & Data Fusion, and the classification process. Firstly: The data preparation process is performed for splitting the input video file to separate the audio and video files, then performing word boundary segmentation to obtain isolated word audio and video files. Secondly: Performing feature extraction for the audio file using MFCC, while using either DCT [2], blocked DCT or LBP with HOG to extract visual features from the mouth region image. Thirdly: Appling dimension reduction methods like Zigzag scanning, PCA, LDA, t-SNE to reduce the visual feature size. Fourthly: Interpolation and integration are then applied to audio and visual extracted features and performing data fusion to obtain a combined feature vector. Finally: a comparison between different classification methods to obtain a higher recognition rate is performed.

A. Data preparation stage

This subsection introduces the important preparation steps required to format the dataset for the feature extraction step and after that for the recognition process.

1) *Extractions of Alignment Audio file from video file:* The existing audio files for the GRID dataset suffer from incorrect word alignment with the accompanying video file [20] so that the FFMPEG command is used to mono channel extraction of the audio file.

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

2) *Word boundary segmentation:* The main step in the data preparation stage is to perform a word boundary separation process to obtain a separated file for each word in the dataset [25]. Based on the information in the alignment file for the GRID dataset the separation process has occurred with the aid of MATLAB [26] program.

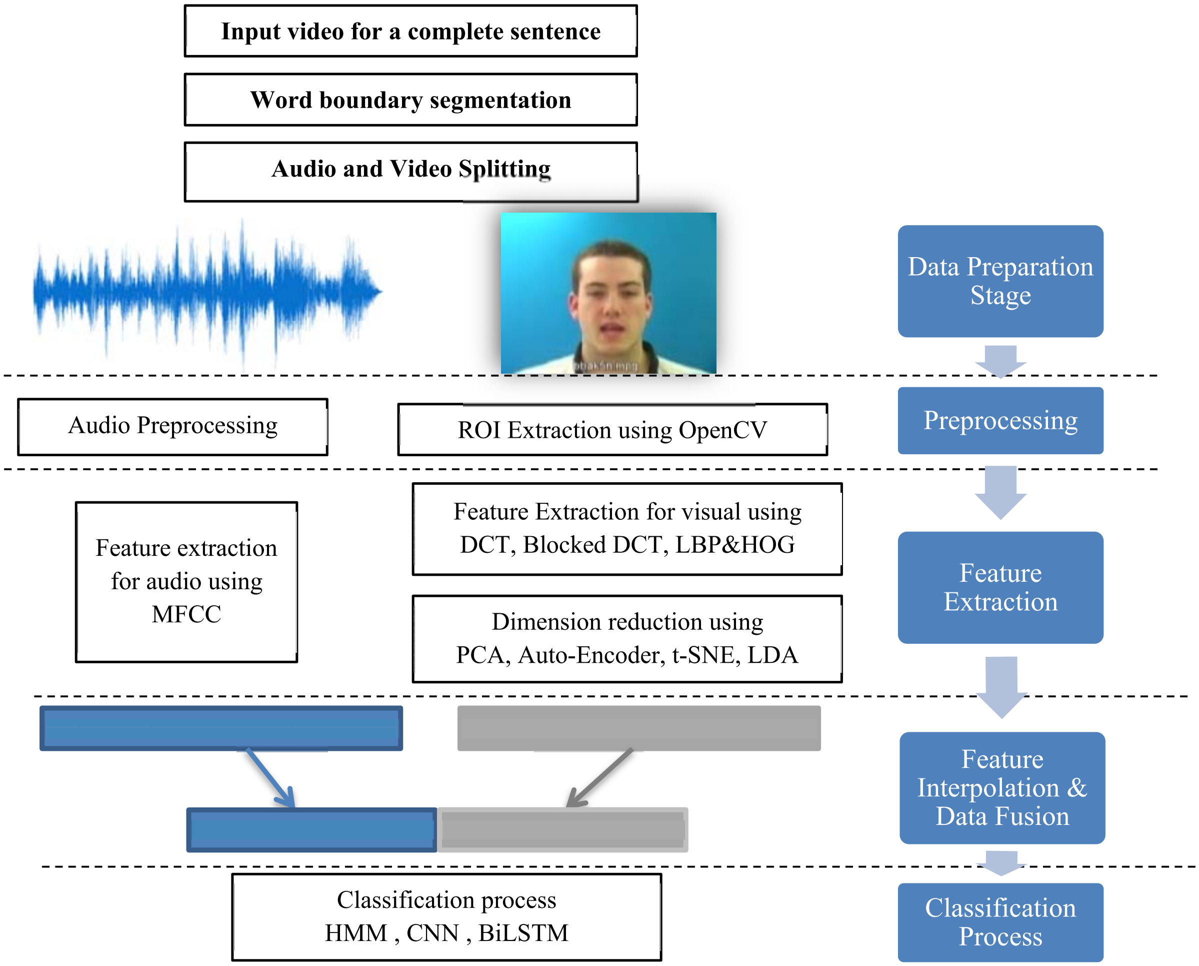
Input video for a complete sentence
Word boundary segmentation
Audio and Video Splitting
Data Preparation Stage
Audio Preprocessing
ROI Extraction using OpenCV
Preprocessing
Feature extraction for audio using MFCC
Feature Extraction for visual using DCT, Blocked DCT, LBP&HOG
Dimension reduction using PCA, Auto-Encoder, t-SNE, LDA
Feature Extraction
Feature Interpolation & Data Fusion
Classification process HMM , CNN , BiLSTM
Classification Process

Figure 1: Pipeline of the proposed AV-ASR model, example for input image taken from GRID dataset [27]

B. Visual Front-End

To take benefits of the visual speech the signal in the speech recognition process, there are important pre-processing steps that must first be performed to obtain the informative speech features.

1) Pre-Processing for Visual signal: The input video is divided into images in the framing stage, this image considers input for the ROI detection stage. Visual speech recognition focuses on the lips of the speaker and removes the background information and other face parts that are not valuable in the speech recognition process. Viola-Jones algorithm [28] is used to detect the mouth region in aid of OpenCV [29] Face-detector module.

2) Feature Extraction for Visual signal: There are three categories of visual feature extraction techniques: pixel-based, shape-based, and hybrid technique [15]. Pixel-based feature extraction technique assumes that all pixel values in ROI are informative for the speech signal [30], while in shape-based, the contour of the speaker's lips contains the informative value for the speech signal. The hybrid method assumes that both pixel and contour values are informative. Extraction of the visual feature based on appearance-based methods used image transformation, like DCT, Discrete Wavelet Transform (DWT), and LDA. In this research, we test using different visual features extraction methods like DCT, blocked DCT followed by PCA and HOG with LBP features.

- **DCT** is used in this research due to its good performance in the previously discussed AV-ASR systems [31] for visual feature extraction either for visual speech or face recognition process. DCT is very popular to use in feature extraction from ROI because of its fast computation using Fast Cosine Transform (FCT) algorithm [32]. After obtaining the DCT for the input image zigzag scanning is used in the coefficients selection stage from the upper left corner of the DCT matrix.
- **Blocked DCT followed by PCA:** Feature extraction using block-based DCT involves dividing the image into blocks of uniform size and isolating the most relevant features of each block. For each block 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 selecting the coefficients according to specific criteria, such as 'low frequency' [20]. Figure 2 explains the used strategy to extract the visual feature using blocked DCT with PCA inspired by the cascade strategy by [33].

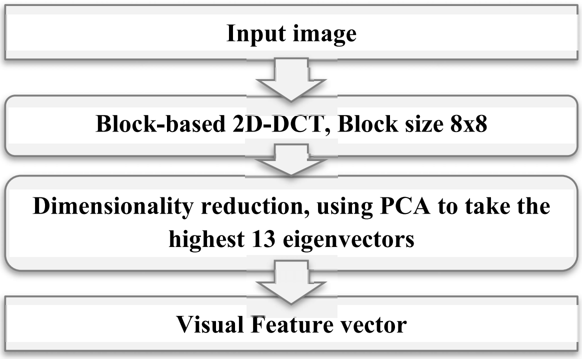
Input image
Block-based 2D-DCT, Block size 8x8
Dimensionality reduction, using PCA to take the highest 13 eigenvectors
Visual Feature vector

Figure 2: Visual feature extraction steps using blocked DCT

- **HOGs and LBPs:** HOG [11] and LBP have proven to be a good descriptor for object recognition in general and face recognition in particular and can also be used in visual speech recognition. In [33] the authors successfully applied HOG descriptors to the problem of face recognition. To compensate for errors in mouth feature detection due to occlusions, pose and illumination changes, we propose to extract HOG descriptors from a regular grid. Then, combining HOG descriptors with the LBP ones allows capturing important structures for mouth recognition. Finally, we use PCA to identify the necessity of performing feature selection to remove redundant and irrelevant features to make the classification process less prone to overfitting [35].

C. Audio Front-End

To make the audio signal suitable for the feature extraction stage there are important pre-processing steps: 1) Framing, where the signal is divided into small pieces called frames as the speech signal is assumed to be stationary for a small interval of time [20]. 2) Frame overlapping, to make sure not to lose the continuity of the speech signal and recovers some (or even all) of the lost signal information. 3) Frame scaling, done by multiplying by Hamming window. After that data is suitable for the feature extraction process, in this paper MFCC [20] is used to extract the features from the audio signal by using HMM Toolkit (HTK) [36].

D. Audio-Visual Features Fusion

The fusion of audio and visual features is used to take advantage of the visual and audio features in the recognition process. The integration of the audio and visual features is called early integration (feature fusion), which is used to obtain a combined feature vector that is passed to the classifier to perform the recognition process. In this paper, early integration is performed by concatenating the audio and visual features. Before performing the concatenation between audio-visual features, we need to make sure that both must have the same feature extraction rates. The feature extraction rate for the audio signal is greater than the feature extraction rate for the visual ones which has the same rate as the input video files. Linear interpolation is used to up-sample the frame rate of the visual signal to make the concatenation between the audio and visual features possible.

E. Classification

HMMs proved to be the most powerful classification method for speech recognition for decades. Although the achievements of HMM in the classification process, deep learning gives more accurate results due to its self-learning mechanism [37]. In this paper, we compare the obtained results from the proposed model using the BiLSTM classifier with HMM, and CNN to get the optimal accuracy for our recognition system.

The adopted CNN architecture consists of ten layers. The first layer is the input feature matrix. The middle layers are four convolution layers, each followed by a max-pooling layer. The last layer is one fully-connected layer to extract the final features. A detailed illustration of the proposed network architecture is shown in figure 3. The rectified linear unit (RELU) is used as an activation function in both convolution layers since it is linear, drivable, and has a simple implementation. The max-pooling operation down-samples the extracted features from the convolution layer. For the other three convolution layers, the same configurations of the first convolution layer are used. Experiments were done on Adam optimizer and loss function categorical cross-entropy.

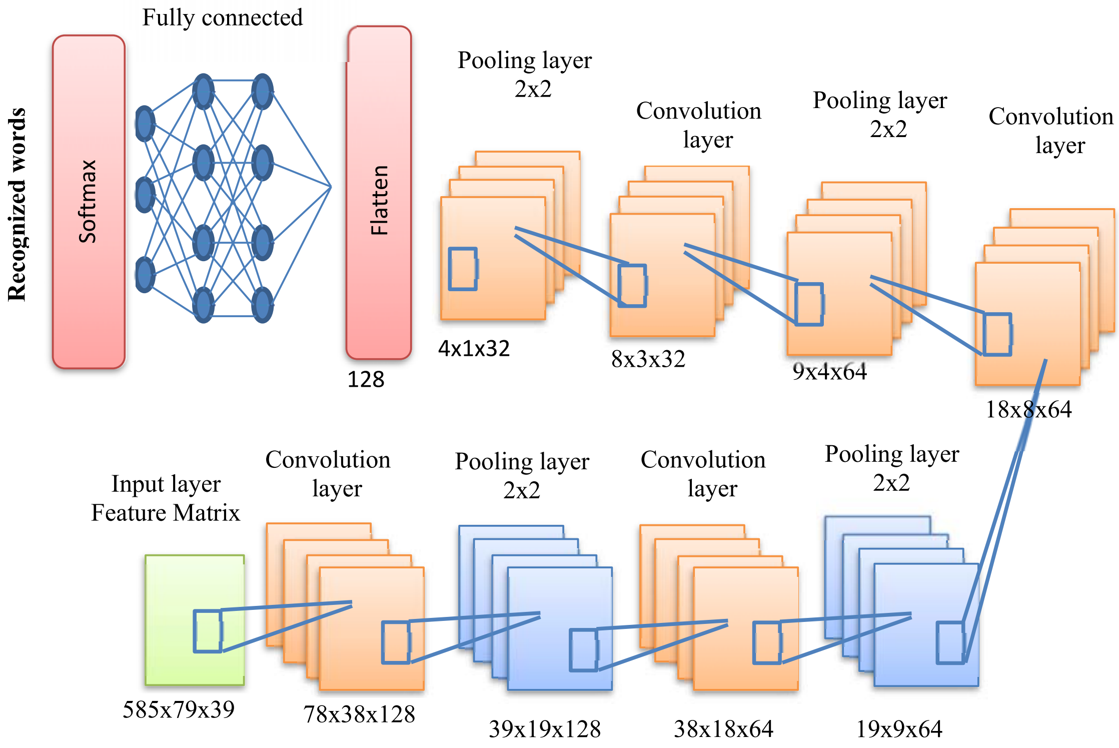

Figure 3: The proposed CNN architecture model

RNN classification method takes advantage of look at the current and the previous frame especially at the word boundary to obtain a robust classification process compared to the other well-known classification methods. In our developed model, Bidirectional LSTM [38] is used to avoid the problem of gradient vanishing and exploding problem [39] that faces the RNN classification. In our proposed model shown in figure 4 the extracted features are used at the input layer, then two copies of the hidden layer of LSTM are created, one fit in the input sequences as-is and one on an opposed copy of the input sequence. The output values from these LSTMs will be concatenated. Each one of the two hidden layers will have 100 memory units (smart neurons) and the output layer will be a fully connected layer that outputs one value per timestep. A softmax activation function is used on the output to predict the isolated word. Because of its great enhancement proved by the previous system, we decided to use BiLSTM in comparison to CNN and HMM for the classification process in our model. All the used parameters of the BiLSTM model are shown in table 1.

TABLE 1
PARAMETERS FOR THE BILSTM MODEL

Parameters	Values
No. of memory units	100
BiLSTM layer activation	sigmoid
Dense layer activation	softmax
Optimizer	Adam
Loss	categorical cross-entropy

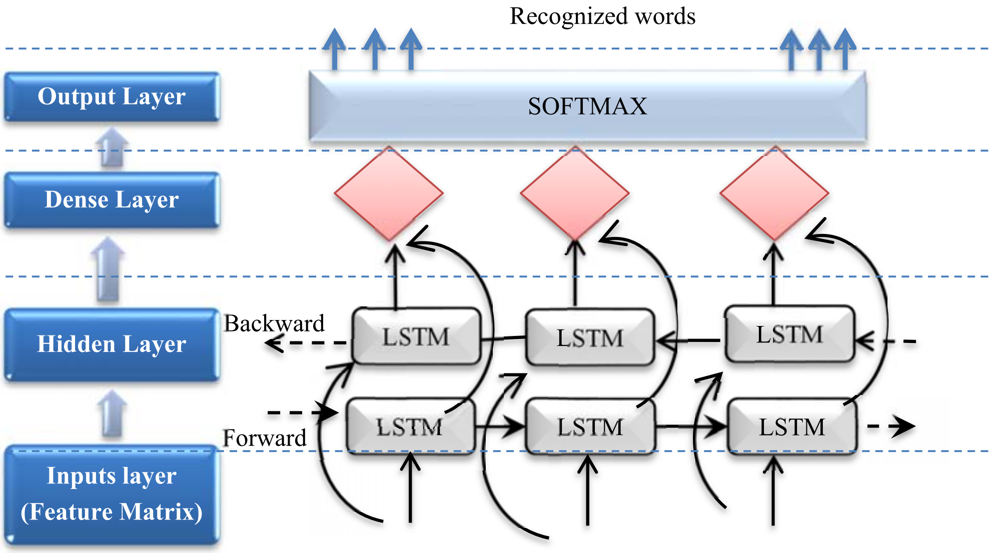

Figure 4: BiLSTM proposed classifier structure

HMM toolkit (HTK) [36] is used for training and testing the HMM model. Each word has separated the HMM model, for the GRID dataset there are 51 HMM models and 26 HMM models for the AVLetters dataset. The performance of the developed system has been evaluated with 5 states of HMM with various numbers of Gaussian Mixtures Model (GMM) which is gradually increased from 2 to 128 mixtures to choose the optimal number of mixtures which give higher recognition accuracy.

3 PROCEDURE AND SYSTEM MODEL

This section presents the datasets used for training and validating the proposed model and explains the grammar file that follows the sequence of the spoken sentences.

The proposed model is evaluated in large audio-visual datasets like GRID which is a continuous audio-visual speech corpus in English. It consists of 34 speakers, each spoken sentence consists of 6 different words and each speaker says 1000 sentences. The grammar for the isolated word recognition process for GRID is shown in figure 5.

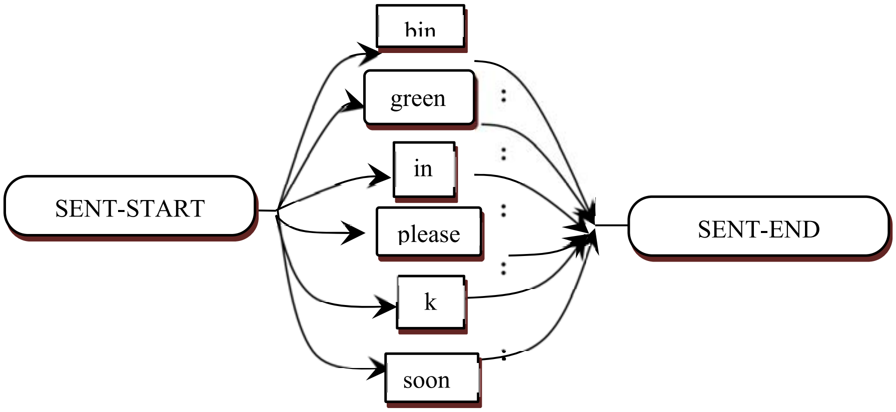

FIGURE 5: GRAMMAR FOR WORD RECOGNITION, GRID DATASET

Also, to perform speaker-independent experiments the proposed model is tested in a small benchmark dataset like the AVletters [2], where ten different speakers repeat the isolated letters A-Z three times, a total of 10x3x26=780 video files. The grammar of the AVletters corpus is shown in figure 6.

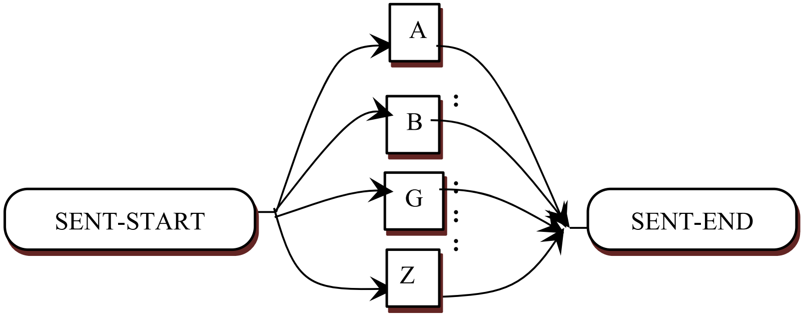

Figure 6: Grammar for word recognition, AVletters dataset

4 Results

The results in this section are obtained by evaluating the proposed model on two well-known audio-visual speech datasets GRID and AVletters to show the effectiveness of CNN, BiLSTM and HMM in AV-ASR. The recognition accuracy metric is used to measure the algorithm's performance in an interpretable way. Accuracy of a model is usually determined after the model parameters and is calculated in the form of a percentage. It is the measure of how accurate your model's prediction is compared to the true data. The recognition accuracy for HMM is calculated by:

$$\%\text{Accuracy} = \frac{N-D-S-I}{N} \times 100 = \frac{H-I}{N} \times 100 \quad [20]$$

where N is the total number of labels in the reference transcriptions, S is the number of substitution errors, D is the deletion errors D and I is the insertion errors.

In case of CNN and BiLSTM, there is a loss function which is used to optimize a machine learning algorithm. The loss is calculated by training and validation and its interpretation is based on how well the model is doing in these two sets. It is the sum of errors made for each example in training or validation sets. Loss value implies how poorly or well a model behaves after each iteration of optimization.

A. AVletters Results:

We evaluate the proposed model firstly on the AVletters benchmark dataset. MFCC is used to extract the audio features with a feature vector size of 26, and DCT is used to extract the visual features with a feature vector size of 13. The recognition accuracies of using HMM are shown in table 2 with mono-phone or tri-phone or different Gaussian mixtures where mix2 to mix128 stands for mixtures from 2 to 128. Table 3 shows the results of using two major DNN architectures (CNN and BiLSTM) in the classification process, the gray cells in these tables give the highest recognition values. Figure 7 compares the results obtained by using different classifier methods and feature types to identify the best model which gives the highest recognition accuracy.

TABLE 2
% ACCURACY RESULT OF HMM FOR AVLETTERS USING (V) VIDEO-ONLY, (A) AUDIO-ONLY, AND (AV) AUDIO-VISUAL FEATURES WITH DIFFERENT GAUSSIAN

No. of Mix Feat Type	**Mono-phone**	**Tri-phone**	**mix2**	**mix4**	**mix8**	**mix16**	**mix32**	**mix64**	**mix128**
A	76.92%	76.41%	80%	82.56%	81.54%	67.18%	27.69%	15.38%	9.74%
AV	68.72%	79.49%	77.95%	79.49%	80%	65.64%	31.79%	14.36%	9.23%
V	27.69%	40%	52.31%	63.08%	77.95%	81.03%	74.87%	69.74%	70.26%

TABLE 3
% ACCURACY AND LOSS RESULT OF CNN AND BILSTM FOR AVLETTERS USING VIDEO-ONLY, AUDIO-ONLY, AND AUDIO-VISUAL

DNN architecture	**BiLSTM**		**CNN**	
Feature Type	**Accuracy**	**Loss**	**Accuracy**	**Loss**
A	86.15%	0.79%	63.5897%	2.50%
V	85.13%	0.81%	**90.256%**	0.53%
AV	**93.33%**	**0.38%**	84.1026 %	1.88%

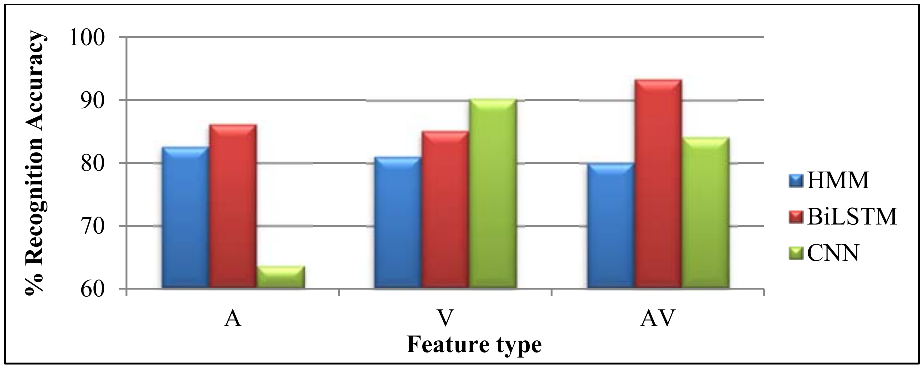

Figure 7: %Recognition accuracy results for A stands for audio-only, V stands for video-only, and AV stands for audio-visual features and different classifiers HMM, CNN, and BiLSTM

Based on the obtained results, we found that:

- Using BiLSTM with an early integrated audio-visual feature gives enhancement over audio-only by 8.33% and decreases the loss value by 45.57%.
- Using CNN with early integrated audio-visual features gives enhancement over audio-only by 32.3% and decreases the loss by 24.8%.
- In the case of HMM, Increasing the Gaussian mixtures in HMM enhances the recognition accuracy until it reaches 8 mixture after that it decreases. The best result is 82.56% with 4 mixtures in audio-only, 81.03% for video-only with 16 mixtures, and 80% for audio-video with 8 mixtures.

The confusion matrix for the BiLSTM AV-ASR model is shown in figure 8 for the AVletters database, it explains the relationship between the true and recognized words where the Greyscale level indicates the density of matching between them.

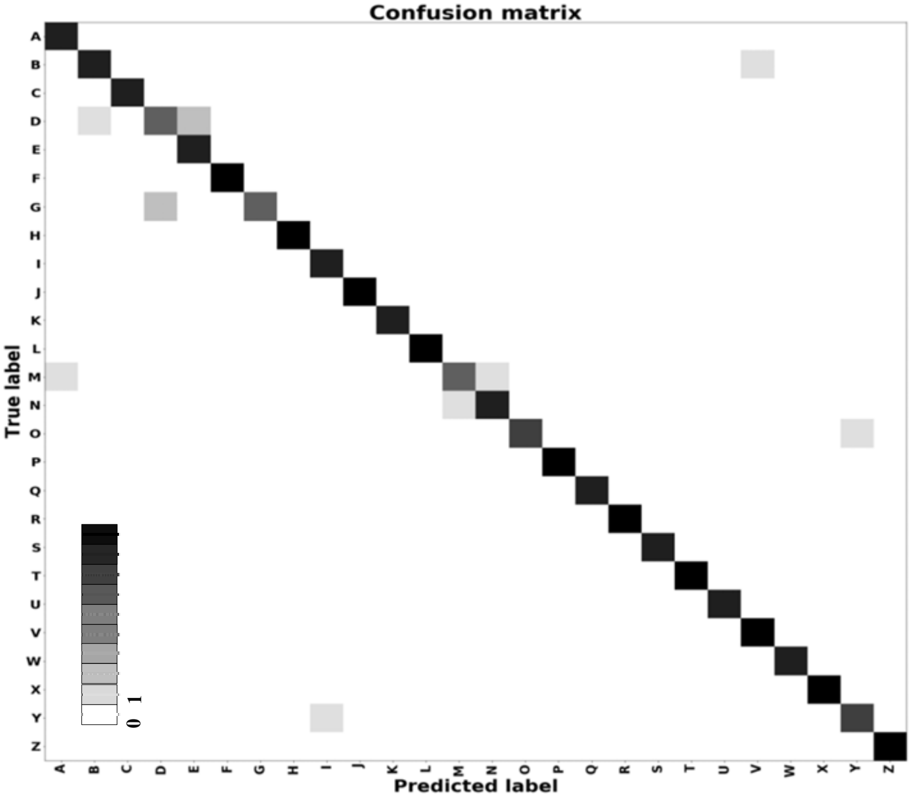

Figure 8: AVletters confusion matrix with audio-visual feature and BiLSTM classifier

The best recognition accuracy is 93.33% when using BiLSTM with early integrated audio-visual feature and enhancement form audio-only up to 8.33 %, which proved that our proposed model gives better recognition accuracy than that obtained in [40] which gives an accuracy of 87.7% for audio-visual using the same dataset as shown in table 4. DCT is used for the input image to extract the main important features then selecting the main important features using zigzag scanning (minimized numbers of features) then feeding these features to the BiLSTM classifier to perform video-only speech recognition.

TABLE 4
THE BEST RECOGNITION ACCURACY OBTAINED BY BILSTM-AVASR FOR AUDIO-ONLY, VIDEO-ONLY, AND AUDIO-VISUAL COMPARED TO THE PREVIOUS RECOGNITION ACCURACY OBTAINED BY USING THE MODEL IN[40].

	Model in [40]	**Proposed model**
Audio-only	75.6%	**86.15%**
Visual-only	64.4%	**85.13%**
Audio-visual	87.7%	**93.33%**

B.GRID Results:

This subsection presents the results of using the GRID dataset in testing the proposed model. The experiments performed in GRID dataset are divided into three stages audio-only with different feature vector size, visual-only with different feature extraction and dimension reduction techniques, and audio-visual speech recognition to ensure the improvement of the combination of audio-visual features to the recognition accuracy. To precisely compare our results to compare the performance of the model to results obtained in [41], we initially performed our experiments on speaker number four (S4, female) from the Grid database as done there.

1) Visual-only Speech Recognition: Table 5 shows a comparison between the performance of using different classification techniques HMMv, CNNv and BiLSTMv (subscript v for visual only) when utilizing DCT or Blocked DCT or HOG+LBP feature extraction in the visual front-ends. The result obtained demonstrates that the performance of the VSR system based on the BiLSTMv model is better than the HMMv and CNNv model, especially when fed with DCT visual feature. The utilization of DCT in BiLSTMv model termed as DCT- BiLSTMv significantly outperforms the traditional DCT-HMMv model by increasing the accuracy from 52.47% to 78.87% with about 50.3% relative improvement. The results show that the deep DCT-BiLSTMv VSR model outperforms the other eight VSR models illustrated in table 5.

TABLE 5
% ACCURACY RESULTS OF THE HMM$_V$, CNN$_V$, AND BILSTM$_V$ MODELS WITH VIDEO-ONLY (V$_D$) DCT OR (V$_{BD}$) BLOCKED DCT OR (V$_{HLBP}$) HOG+LBP AS INPUT FEATURES

	HMM$_v$	**BiLSTM$_v$**	**CNN$_v$**
V$_D$	52.47 mix16	***78.87***	75.6
V$_{BD}$	23.33 mix64	**46.8**	44.26667
V$_{HLBP}$	27.07 mix64	**42.4**	40.53333

2) Audio-only Speech Recognition: Table 6 shows a comparison between the performance of the HMMa with Gaussian Mixtures from 2 to 128, CNNa and BiLSTMa ASR models after utilizing MFCC_0 (A13 stands for size of audio feature vector 13) or MFCC_D_A_0 (A39 stands for size of audio feature vector 39) features in the audio front-ends for clean data and after adding babble noise with different SNR. It demonstrates that the accuracy of the speech recognition system based on the BiLSTMa model is better than the HMMa and CNNa model, especially when fed with A39. The utilization of BiLSTMa model significantly outperforms the traditional HMMa model by increasing the accuracy from 89.51% to 98.2% with about 9.7% relative improvement. Increasing the Gaussian mixtures for HMMa model enhances the recognition accuracy till it reaches mix8 after that it decreases, the improvement of A39-BiLSTMa becomes more obvious when adding the noise signal. The results show that the deep A39-BiLSTMa ASR model outperforms the other ASR models.

TABLE 6
% ACCURACY RESULTS OF THE HMMA, CNNA AND BILSTMA MODELS FOR AUDIO-ONLY A13 OR A39 AS INPUT FEATURES, WITH DIFFERENT SNR

		0db	**5db**	**10db**	**15db**	**20db**	**clean**
HMMa	**A13**	86.24 Mix16	86.78 Mix8	87.6 Mix8	87.26 Mix8	87.94 Mix8	89.51 Mix8
	A39	92.27	93.53	95.07	95.27	95.4	95.93
BiLSTMa	**A13**	94.13	95	95.47	97.73	98.13	**98.2**
	A39	**95.73**	**97**	**97.23**	**97.2**	**97.36**	97.47
CNNa	**A13**	93.53	95.04	95.13	95.23	95.33	95.53
	A39	92.27	93.53	95.07	95.27	95.4	95.93

3) Audio-Visual Speech Recognition: Figures 9 and 10 show a comparison between the accuracy of HMMav, CNNav and BiLSTMav AV-ASR models with different SNR after utilizing DCT or Blocked DCT or HOG+LBP features in the visual front-ends and MFCC_0 or MFCC_D_A_0 in the audio front-ends, and using Early Integration (EI) scheme to get the combined feature vector. In figure 9 AV13D, AV13BD, and AV13HLbp stands for early integrated MFCC_0 with DCT, blocked DCT and HOG+LBP features respectively, while AV39D, AV39BD, and AV39HLbp stand for early integrated MFCC_D_A_0 with DCT, blocked DCT and HOG+LBP features respectively. It demonstrates that the accuracy of the audio-visual speech recognition system based on the BiLSTMav model is better than the HMMav and CNNav model, especially when fed with DCT with MFCC_D_A_0. The utilization of DCT+ MFCC_D_A_0 in the BiLSTMav model, termed as AV39D-BiLSTMav, significantly outperforms the traditional AV39D-HMMav model by increasing the accuracy from 94.4% to 99.13% with about 5.01% relative improvement. The results show that the deep AV39D-BiLSTMav AV-ASR model outperforms the other AV-ASR models.

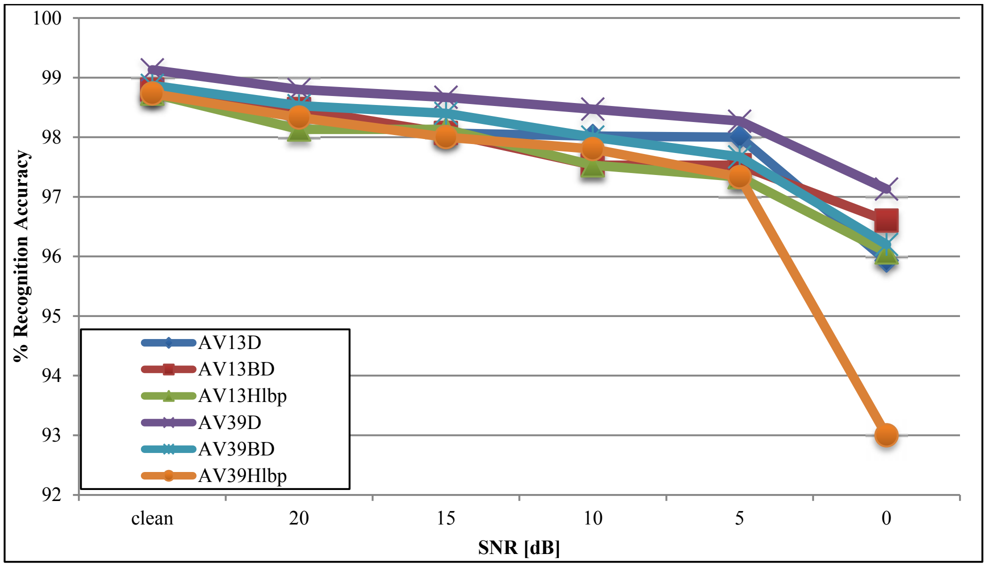

Figure 9: Comparison between the performance of different combinations for audio and visual features with the BiLSTMav classifier.

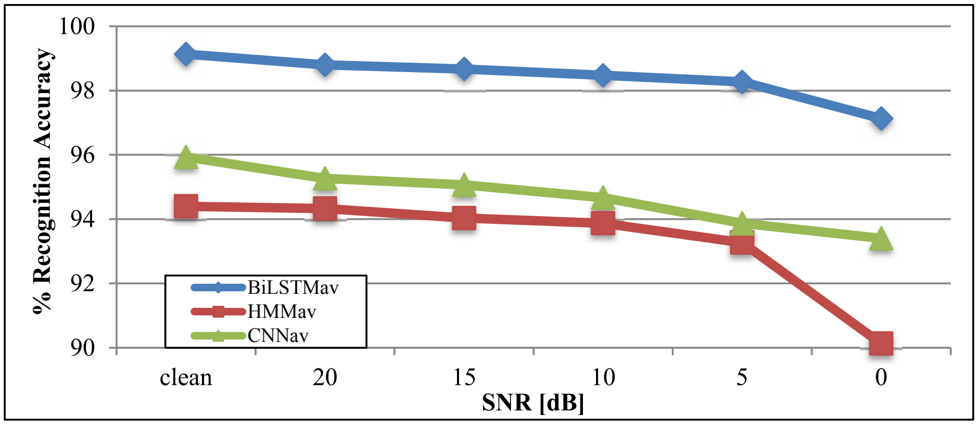

Figure 10: Performance comparison of deep BiLSTMav, HMMav, and CNNav models using MCC_D_A_0 as audio features and DCT as visual features at different SNR levels.

4) Result for different dimension reduction methods: To reduce the visual feature vector size without losing the main feature. There is a lot of dimension reduction techniques that can be used to test their effect on recognition accuracy. The result of using different dimension reduction techniques like PCA, auto-encoder, LDA, t-SNE is shown in table 7, where AV_DCT_t-SNE stands for early integrated the visual features obtained by DCT with audio feature and using t-SNE in the dimension reduction stage, AV_DCT_PCA using PCA, AV_DCT_LDA using LDA, AV_DCT_auto-encoder using auto-encoder. Matlab Toolbox for Dimensionality Reduction [42] is used to perform different dimension reduction techniques. From the obtained results it is shown that using DCT with PCA gives the best recognition accuracy when compared to other techniques.

TABLE 7
%RECOGNITION ACCURACY FOR DIFFERENT DIMENSION REDUCTION TECHNIQUES WHEN USING HMM FOR CLASSIFICATION

		Mono-phone	Tri-phone	mix2	mix4	mix8	mix16	mix32	mix64	mix128
AV_DCT_t-SNE	SENT	89.47	97.73	97.93	98.4	98.53	98	90.8	77.13	63.4
	WORD	51.2	89.6	90.8	93.2	94	90.4	59.6	20.4	1.2
AV_DCT_PCA	SENT	91.47	98.47	98.4	98.8	98.93	98.33	92.27	78.6	64.87
	WORD	56	90.8	90.4	92.8	93.6	90.4	63.2	2	24.8
AV_DCT_LDA	SENT	90.27	98.2	98.47	98.6	98.8	98.13	91.2	76.53	60.4
	WORD	52	89.6	91.2	91.6	93.2	89.6	62.4	22.8	3.2
AV_DCT_ Auto-encoder	SENT	86.2	97.4	97.93	98.33	98.53	96.6	88.73	74.47	62.4
	WORD	42.8	86.8	88.8	90.8	92.8	83.2	51.2	18.4	2.4

Our proposed model achieves improvement in the recognition accuracy. For audio-visual features enhancement up to 24% and for audio-only up to 18.8% for speaker 4 when using the BiLSTM classifier. The obtained result is better than the recognition accuracy achieved by Ephrat [41] for the same speaker.

We also perform the experiments on speakers no. 6, 11, and 12 (one female and two males) to ensure the robustness of our model. From the obtained results illustrated in table 8 and table 9, we can conclude the following:

- Using HMMav, HMM classifier with early integrated DCT+MFCC audio-visual feature improved the recognition accuracy over audio-only up to 3.35%, 2.27%, and 1.45% in a clean environment, while after adding babble noise with SNR 5db the improvement is 3.89%, 8.73%, and 1.71% for speaker 6, 11 and 12 respectively.
- Using CNNav, CNN classifier with an early integrated audio-visual feature produces better recognition accuracy than using an audio-only feature, in the clean environment and after adding babble noise with 5db SNR for the three speakers.
- When using BiLSTMav, BiLSTM classifier introduces the best accuracy that occurred over the HMM and CNN classifiers when using early integrated audio-visual MFCC_D_A_0+DCT feature in clean and noisy environment for the three speakers.
- The loss value decreases when adding the visual features to audio features with feature vector size of 39 in both clean and noisy audio signals and when using either CNN or BiLSTM classification methods as shown in table 9.

TABLE 8
% ACCURACY RESULTS FOR SPEAKERS 6, 11, AND 12 FOR GRID DATASET

Classification methods	Spk no.	Clean					Noise 5db				
		A13	A39	V13	AV13	AV39	A13	A39	V13	AV13	AV39
BiLSTM	S6	98.90%	99.10%	93.40%	99.60%	99.70%	97.60%	98.50%	93.40%	99.50%	99.50%
	S11	96.30%	96.70%	90.30%	97.80%	98.20%	94.10%	94.90%	90.30%	96.10%	96.90%
	S12	99.30%	99%	93.90%	99.70%	99.80%	98.50%	98.60%	93.90%	99.50%	99.60%
CNN	S6	97%	97%	97.40%	98.50%	98.70%	95.50%	96.80%	97.40%	97.50%	96.40%
	S11	91.20%	91.70%	96.50%	95.80%	96.90%	88%	88.70%	96.50%	95.40%	96.50%
	S12	96.20%	96.90%	97.10%	98.70%	98.90%	94.90%	95.10%	97.10%	98.80%	98.80%
HMM	S6	94.54 mix16	94.57 mix16	94.34 mix16	97.71 mix16	98% mix16	91.12 mix16	92.4% mix16	94.34 mix16	95% mix16	96% mix16
	S11	86.7% mix8	87.9% mix8	80.3% mix8	88.7% mix8	89.9% mix8	74.4% mix8	79.7% mix8	80.3% mix8	80.9% mix8	79.1% mix8
	S12	95.2% mix8	95.9% mix8	84.1% mix8	95.5% mix8	97.3% mix8	91.7% mix4	93.3% mix4	84.1% mix8	88.5% mix4	94.9% mix4

TABLE 9
LOSS VALUE RESULTS FOR SPEAKERS 6, 11, AND 12 FOR GRID DATASET

	SNR	Feature type	S6	S11	S12
BiLSTM	Clean	A13	0.03%	0.11%	0.02%
		A39	0.02%	0.10%	0.02%
		V13	0.20%	0.28%	0.18%
		AV13	0.01%	0.06%	0.01%
		AV39	0.01%	0.05%	0.01%
	5db	A13	0.07%	0.17%	0.05%
		A39	0.04%	0.17%	0.04%
		AV13	0.02%	0.10%	0.01%
		AV39	0.02%	0.09%	0.01%
CNN	Clean	A13	0.09%	0.29%	0.12%
		A39	0.09%	0.31%	0.13%
		V13	0.08%	0.11%	0.09%
		AV13	0.05%	0.14%	0.04%
		AV39	0.05%	0.08%	0.04%
	5db	A13	0.15%	0.35%	0.16%
		A39	0.25%	0.41%	0.19%
		AV13	0.07%	0.10%	0.04%
		AV39	0.09%	0.10%	0.04%

Figure 11 gives a comparison between the results obtained when using BiLSTM, CNN, and HMM in the classification process with feature type either audio-only or video-only or early integrated audio-visual feature for speaker 6. The utilization of AV39D-BiLSTMav significantly outperforms the traditional AV39D-HMMav model by increasing the accuracy from 95.2% to 99.3% with about 5.01% relative improvement. The results show that the deep AV39D-BiLSTMav AV-ASR model outperforms the other AV-ASR models with other feature types.

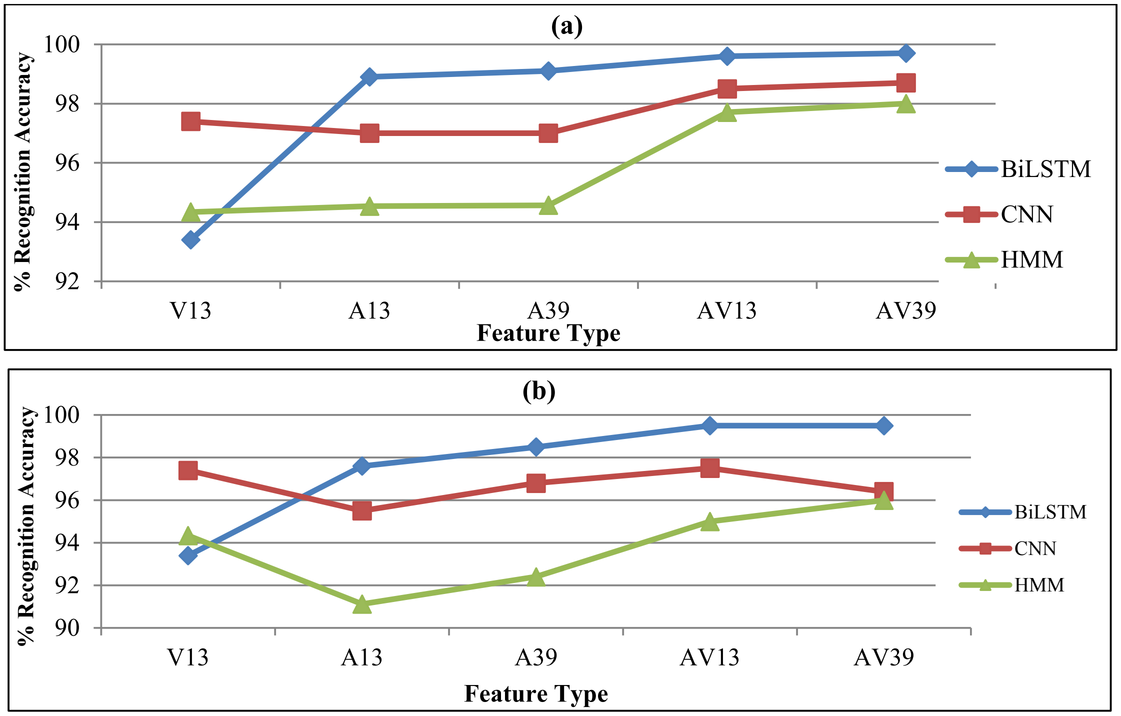

Figure 11:Result of using different classification techniques with audio-only feature MFCC_0(A13) , MFCC_D_A_0(A39) , visual-only(V13) and audio-visual feature MFCC_0+V13(AV13) , MFCC_D_A_0+V13 (AV13) in (a) clean,(b)noisy system.

The confusion matrix for speaker 12 of the GRID database is introduced in figure 12, using the BiLSTM AV-ASR model after early integration of the visual feature obtained by DCT and audio feature MFCC obtained with size 39. The reference labels are represented in rows, and classification postulate is represented in columns.

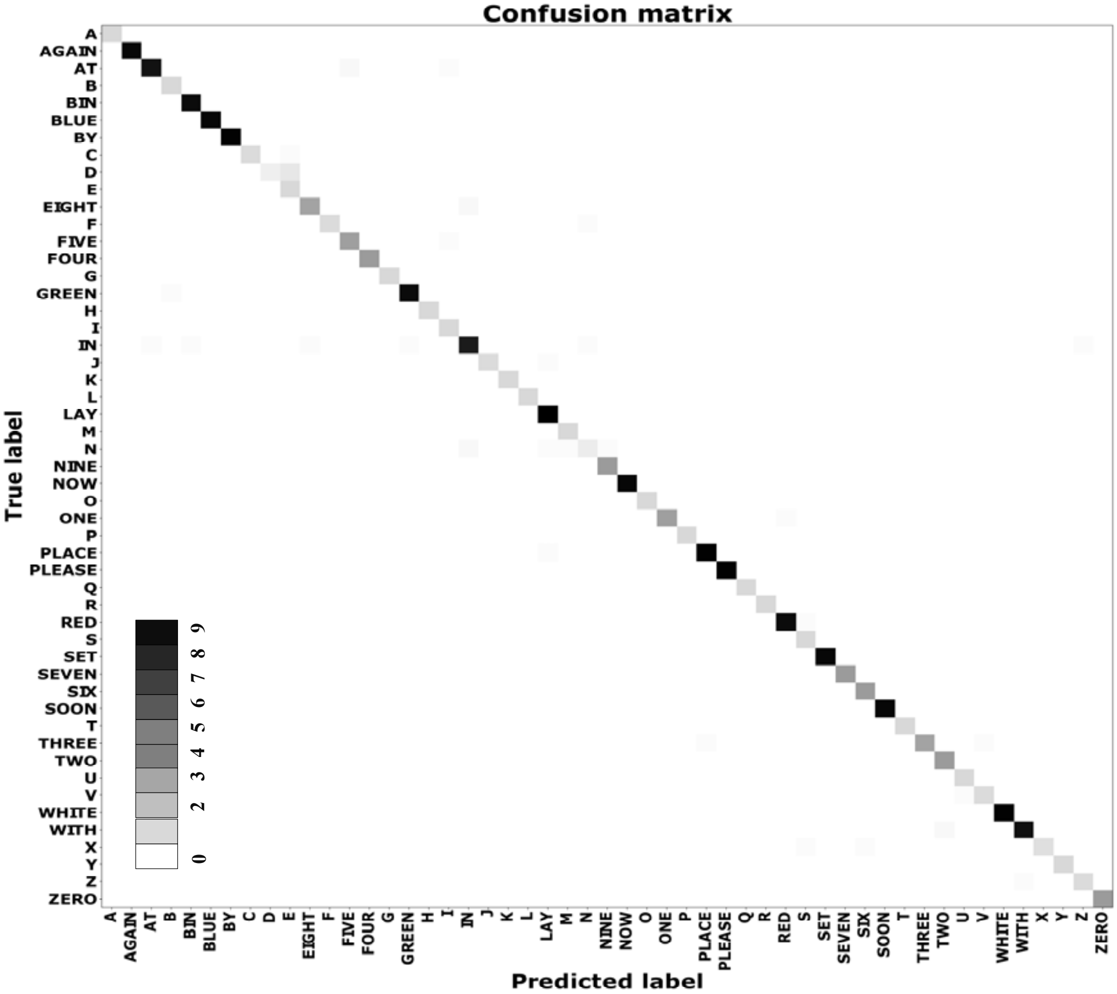

Figure 12: Isolated Word recognition confusion matrix for Speaker 12 of GRID dataset AV39-BiLSTM model

5 Data availability

The datasets analyzed during this research are introduced in [27] and [2]. These datasets are available online for research topics from the following public domain resources:

- AVletters dataset available online: http://www2.cmp.uea.ac.uk/~bjt/avletters/
- GRID dataset available online: http://spandh.dcs.shef.ac.uk/gridcorpus/

6 Conclusion

The work in this paper introduces a novel AV-ASR model that consists of two main stages: feature extraction stage, and classification stage. The feature extraction stage is divided into two sub-stages for audio and visual signal, MFCC is used to extract the most effective feature from the audio signal with vector sizes either 13 or 39. According to the visual feature, different visual feature extraction methods are tested in this paper like DCT or Blocked DCT or HOG+LBP are used to extract the effective features from the lip region. Different dimension reduction techniques are used to select the most important visual components from the extracted visual feature using either PCA or t-SNE or auto-encoder or LDA. Our experimental results demonstrated that extracting the visual features using DCT with zigzag scanning is the best visual feature extraction and dimension reduction technique. The results of applying different classification techniques like BiLSTM, CNN and traditional classifier HMM, BiLSTM proved to be the most effective classifier to obtain reliable and noise-robust speech recognition system. Evaluating the proposed model on two multi-speakers audio-visual datasets AVletters and GRID with speaker-dependent and independent experiments ensure that the performance enhancement of the proposed model, even with the used simple multimodal integration mechanism between audio and visual features. The proposed model with combining DCT visual feature and MFCC audio feature and using the BiLSTM classifier gives a great enhancement in the recognition accuracy and decreasing the loss value for both clean and noisy environments than using audio-only features.

References

[1] E. E. El Maghraby, A. M. Gody, and M. H. Farouk, "Speech Recognition Using Historian Multimodal Approach," *Egypt. J. Lang. Eng.*, vol. 6, no. 2, pp. 44–58, 2019.

[2] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 2, pp. 198–213, 2002.

[3] G. Zhao, M. Barnard, and M. Pietikäinen, "Lipreading with local spatiotemporal descriptors," *IEEE Trans. Multimed.*, vol. 11, no. 7, pp. 1254–1265, 2009.

[4] A. B. A. Hassanat, "Visual Speech Recognition," *Speech and Language Technologies*, no. 1, pp. 279-303, 2011.

[5] A. Chern, Y. H. Lai, Y. P. Chang, Y. Tsao, R. Y. Chang, and H. W. Chang, "A Smartphone-Based Multi-Functional Hearing Assistive System to Facilitate Speech Recognition in the Classroom," *IEEE Access*, vol. 5, pp. 10339–10351, 2017.

[6] A. El-Solh, A. Cuhadar, and R. A. Goubran, "Evaluation of speech enhancement techniques for speaker identification in noisy environments," *In Ninth International Symposium on Multimedia Workshops (ISMW 2007),* pp. 235–239, IEEE, 2007.

[7] F. Tao and C. Busso, "Lipreading approach for isolated digits recognition under whisper and neutral speech," in *Fifteenth Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2014, pp. 1154–1158.

[8] J. Yu, S. X. Zhang, J. Wu, S. Ghorbani, B. Wu , S. Kang,and D. Yu,. "Audio-visual Recognition of Overlapped speech for the LRS2 dataset.", arXiv:2001.01656, 2020.

[9] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent advances in the automatic recognition of audiovisual speech," *in Proceedings of the IEEE, 2003, vol. 91, no. 9, pp. 1306–1325.*

[10] J. Z. Neti, C., G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, "Audio-visual speech recognition final workshop Final Report," *Cent. Lang. Speech Process. Johns Hopkins Univ. Balt.*, 2000.

[11] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[12] G. Potamianos, H. P. Graf, and E. Cosatto, "An Image transform approach for HMM based automatic lipreading," in *Proceedings 1998 International Conference on Image Processing. ICIP98(Cat. No. 98CB36269). Chicago, IEEE.*, 1998, vol. 3, pp. 173–177.

[13] G. Potamianos, C. Neti, and G. Iyengar, "A Cascade Visual Front End for Speaker Independent," *Int. J. Speech Technol.*, vol. 4, pp. 193–208, 2001.

[14] M. Heckmann, K. Kroschel, C. Savariaux, and F. Berthommier, "DCT-based video features for audio-visual speech recognition," in *Seventh International Conference on Spoken Language Processing*, pp.1925-1928, 2002.

[15] J. He and H. Zhang, "Research on visual speech feature extraction," in *International Conference on Computer Engineering and Technology,* vol. 2, pp. 499–502, 2009.

[16] G. Chollet *et al.*, "Some experiments in audio-visual speech processing," *Lect. Notes Comput. Sci. (including Subser.* Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 4885 LNAI, pp. 28–56, 2007.

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

[18] P. Zhou, W. Yang, W. Chen, Y. Wang & J. Jia, "Modality attention for end-to-end audio-visual speech recognition". *In ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6565-6569). IEEE,2019*

[19] E. D. Petajan, "Automatic Lipreading to Enhance Speech Recognition (Speech Reading)," University of Illinois at Urbana-Champaign, 1985.

[20] E. E. El Maghraby, A. M. Gody, and M. H. Farouk, "Enhancing quality and accuracy of speech recognition system by using multimodal audio-visual speech signal," in *12th International Computer Engineering Conference, ICENCO: Boundless Smart Societies*, pp. 219–229. IEEE, Egypt, 2017.

[21] E. S. Salama, R. A. El-Khoribi, and M. E. Shoman, "Audio-Visual Speech Recognition for People with Speech Disorders," *Int. J. Comput. Appl.*, vol. 96, no. 2, pp. 51–56, 2014.

[22] G. Hinton *et al.*, "Deep Neural Networks for Acoustic Modeling in Speech Recognition," *IEEE Signal Process. Mag.*No. 29, vol. 6, pp. 82–97, 2012.

[23] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Audio-visual speech recognition using deep learning," Applied Intelligence, vol. 42, no. 4, pp. 722–737, 2015.

[24] J. C Hou, S. S.Wang Y. H.Lai, Y. Tsao, H. W.Chang and H.M. Wang. "Audio-visual speech enhancement using multimodal deep convolutional neural networks". IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 2, no. 2, pp.117-128.IEEE, 2018.

[25] H. L. Bear and R. Harvey, "Decoding visemes: Improving machine lip-reading," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2009–2013. *IEEE* 2016.

[26] "MATLAB program." [Online]. Available: http://www.mathworks.com.

[27] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. Acoust. Soc. Am.*, vol. 120, no. 5, pp. 2421–2424, 2006.

[28] O. H. Jensen, "Implementing the Viola-Jones face detection algorithm," Technical University of Denmark, DTU, DK-2800 Kgs, Lyngby, Denmark, 2008.

[29] W. G. Intel Corporation, "OpenCV." [Online]. Available: http://www.opencv.org.

[30] G. Potamianos and P. Scanlon, "Exploiting lower face symmetry in appearance-based automatic speechreading," *Audit. Speech Process.*, pp. 79–84, 2005.

[31] V. Estellers and J. P. Thiran, "Multi-pose lipreading and audio-visual speech recognition," *EURASIP J. Adv. Signal Process.*, vol. 1, pp. 1–23, 2012.

[32] J. Chaloupka, "Extraction of the visual features by discrete cosine transform for audio-visual speech recognition," in *Radioelektronika*, 2005, pp. 467–470.

[33] A. V Nefian and K. Murphy, "Dynamic Bayesian Networks for Audio-Visual Speech Recognition," *EURASIP J. Adv. Signal Process.*, vol. 11, pp. 1274–1288, 2002.

[34] A. Albiol, D. Monzo, A. Martin, J. Sastre, and A. Albiol, "Face recognition using HOG – EBGM," *Pattern Recognit. Lett.*, vol. 29, pp. 1537–1543, 2008.

[35] M. Ghorbani, A. T. Targhi, and M. M. Dehshibi, "HOG and LBP: Towards a robust face recognition system," *10th Int. Conf. Digit. Inf. Manag. ICDIM*, pp. 138–141, 2016.

[36] S. Young *et al.*, *The HTK Book*, vol. 3.4. 2006.

[37] J. Schmidhuber, "Deep Learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.

[38] S. Hochreiter and J. Urgen Schmidhuber, "Long Short term Memory," *Neural Comput.*, vol. 9, no. 8, p. 17351780, 1997.

[39] Yoshua Bengio, Patrice Simard, and Paolo Frasconi, "Learning Long-term Dependencies with Gradient Descent is Difficult," *IEEE Transactions on Neural Network*, vol. 5, no. 2. p. 157, 2014.

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

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

[42] L. Van Der Maaten, "Matlab toolbox for dimensionality reduction," *Belgian/Netherlands Artif. Intell. Conf.*, no. 1, pp. 439–440, 2007.

BIOGRAPHY

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 Systems Department, Faculty of Computers and Information, Fayoum University. Her research interest is in signal processing and computer networks.

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 the 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.

TRANSLATED ABSTRACT

تصميم نظام للتعرف على الاصوات قوي في حالة وجود ضوضاء يعتمد على وسائط سمعية وبصرية للصوت مع تقنيات مختلفة للتعلم العميق

إسلام عيد علي محمد المغربي*[1]، عمرو محمد رفعت جودي*[2]، محمد هشام فاروق**[3]

*قسم هندسة الاتصالات والالكترونيات - كلية الهندسة - جامعة الفيوم

** قسم هندسة الرياضيات والفيزيقا - كلية الهندسة - جامعة القاهرة

[1]eem00@fayoum.edu.eg

[2]amg00@fayoum.edu.eg

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

ملخص

يعد إستخدام الإشارة الصوتية بالإضافه الي حركة الشفاه أحد الحلول الواعدة لتصميم نظام قوي للتعرف على الأصوات ، خاصةً عندما تتأثر الإشارة الصوتية بالضوضاء يمكن في هذه الحالة إستخدام حركة الشفاه للحصول على مزيد من المعلومات لتحسين دقة التعرف على الأصوات في بيئة معرضة للضوضاء ، وذلك لأن حركة الشفاه لا تتأثربالضوضاء الصوتية. تتمثل المرحلة الحاسمة في تصميم نظام قوي للتعرف على الأصوات هي إختيار طريقة تصنيف موثوق بها من ضمن مجموعة كبيرة ومتنوعة من تقنيات التصنيف الحالية. يقوم هذا البحث بتكملة عملنا السابق في تصميم نظام للتعرف على الأصوات المعتمد علي تقنية التصنيف HMM بإستخدام حركة الشفاه بالإضافة إلى الاشارة الصوتية. عملية التصنيف تتم بإستخدام نظام قائم على RNN و CNN لإظهار كيف يمكن لكل منهما تحسين دقة التعرف علي الصوت عند مقارنتهما بتقنية HMM في حالة وجود ضوضاء أم لا. مساهمات هذا البحث ذات شقين: أولاً ، من أجل إختيار أكثر الطرق الفعالة لإستخراج خصائص الصوت من حركة الشفاه والتي توفر تحسينًا فعالًا لنظام التعرف على الأصوات ، لتحقيق ذلك يقدم هذا البحث مقارنة بين الطرق المختلفة لإستخراج خصائص الصوت من حركة الشفاه مثل DCT و DCT وBlocked HOG+LBP , ثم إستخدام تقنيات مختلفة لتقليل حجم الخصائص المستخرجة من حركة الشفاه للعثور على أفضل حجم لها مثل PCA و auto-encoder و LDA و t-SNE . بعد ذلك يتم دمج الخصائص المستخرجة من حركة الشفاه وخصائص الإشارة الصوتية التي يتم إستخراجها بإستخدام MFCC. ثانياً: يتم تنفيذ عملية التصنيف بإستخدام إحدي تقنيات DNN المختلفة مثل BiLSTM ومقارنة النتائج مع CNN وHMM. يتم تقييم النموذج المقترح AV-ASR بإستخدام قواعد بيانات متعددة المتحدثين مثل AVletters و GRID مع قيم SNR مختلفة. يُجري النموذج تجارب لا تعتمد علي متحدث معين بإستخدام قاعدة بيانات AVlettter وتجارب تعتمد علي كل متحدث علي حدى بإستخدام قاعدة بيانات GRID. أظهرت النتائج التجريبية التي تم تقديمها في هذا البحث أن إستخدام الخصائص التي تدمج بين خصائص الصوت التي تم الحصول عليها باستخدام MFCC وخصائص حركة الشفاه التي تم الحصول عليها بإستخدام DCT تظهر دقة عالية في التعرف علي الصوت عند إستخدامها مع تقنية تصنيف BiLSTM مقارنة بطرق أخرى لتقنيات إستخراج وتصنيف الخصائص الصوتية.

الكلمات المفتاحية: AV-ASR ، HMM ، DNN ، AVletters ، GRID ، DCT ،Blocked DCT ، PCA