Research Article

International Journal of Advanced Computer Research, Vol 10(47)
ISSN (Print): 2249-7277   ISSN (Online): 2277-7970
http://dx.doi.org/10.19101/IJACR.2019.940134

Noise robust speech recognition system using multimodal audio-visual approach using different deep learning classification techniques

Eslam E. El Maghraby[1*], Amr M. Gody[2] and Mohamed Hesham Farouk[3]
Assistant Lecturer, Department of Computers and Information, Fayoum University, Egypt[1]
Faculty, Department of Electrical Engineering, Cairo University, Egypt[2]
Professor, Department of Math & Physics, Cairo University, Egypt[3]

Received: 06-November-2019; Revised: 15-January-2020; Accepted: 10-March-2020

Abstract

Multimodal speech recognition is proved to be one of the most promising solutions for designing robust speech recognition system, especially when the audio signal is corrupted 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 signal and the choice of a reliable classification method from a large variety of existing classification techniques. This paper proposes an Audio-Visual Speech Recognition (AV-ASR) system using both audio and visual speech modalities to improve recognition accuracy in a clean and noisy environment. The contributions of this paper are two-folded: The first is the methodology of choosing the visual features by comparing different 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. These features are then early integrated with audio features obtained by Mel frequency Cepstral coefficients (MFCCs) and feed into classification process. The second contribution of this research is the methodology of developing the classification process using deep learning, comparing different deep neural network (DNN) architectures like bidirectional long-short term memory (BiLSTM), and convolution neural network (CNN), with the traditional hidden Markov models (HMM).The effectiveness of the proposed model is demonstrated on two multi-speakers AV-ASR benchmark datasets named AVletters and GRID with different SNR. The model performs speaker-independent experiments in AVlettter dataset and speaker-dependent for the GRID dataset. The experimental results show that early integration between audio feature obtained by a MFCC and visual feature obtained by DCT demonstrate higher recognition accuracy when used with BiLSTM classifier compared to other methods for features extraction and classification techniques. In case of GRID, using integrated audio-visual features achieved highest recognition accuracy of 99.13% and 98.47%, with enhancement up to 9.28% and 12.05% over audio-only for clean and noisy data respectively. For AVletters, the highest recognition accuracy is 93.33% with enhancement up to 8.33% over audio-only. The obtained results show the performance enhancement compared to previously obtain audio-visual recognition accuracies on GRID and AVletters and prove the robustness of our BiLSTM-AV-ASR model when compared with CNN and HMM, because BiLSTM takes into account the sequential characteristics of the speech signal.

Keywords

AV-ASR, DCT, Blocked DCT, PCA, MFCC, HMM, BiLSTM, CNN, AVletters and GRID.

1.Introduction

Speech understanding of human is performed by using audio and visual information e.g. movements of speaker lips and tongue.

Speech is a multimodal signal that depends on audio and visual modalities, so to build a 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 lip movements to identify the spoken words called lipreading. Lipreading can be used in addition to the audio signal to enhance speech recognition

[*] Author for correspondence

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.

The 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 has 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 transform (DCT) [2], discrete wavelet transforms (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 a 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 in speech recognition for noisy environment encourages researchers to use vision in addition to hearing a speech recognition system. Potamianos [12] summarized the main components required to build a robust AV-ASR system.

Choosing the best classification techniques has been a significant research topic for decades, either for visual-only or AV-ASR 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 systems 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 computation 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].

1.1Problem statement

The 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.

There are important tasks in the speech recognition process have up till now to be successfully addressed: 1) Selecting the most effective visual feature extraction method, 2) Integration of the acoustic and visual speech modalities successfully, and 3) Selecting the most effective classification techniques. A lot of machine learning algorithms can be used in the classification process like KNN, SVM, logistic regression etc. These algorithms, learning less while comparing to CNN and RNN that is because of there is no transfer learning does happen in the Machine learning algorithm, while there is transfer learning in deep learning. This enables deep learning to learn more, and also less error will occur. This paper aims to propose a robust and reliable approach to deep

AV-ASR model by using deep learning classification engine long short-term memory bidirectional recurrent neural network (BiLSTM) with early integration (EI) scheme and compared its results to CNN and traditional HMM classifier to ensure the robustness of the proposed speech recognition system.

1.2Paper 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 functional 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

A list of the latest and most relevant algorithms used in building speech recognition system is discussed in this section. The previous work in speech recognition system can be divided into two main stages feature extraction and classification process. In feature extraction stage, in addition to the well-known visual feature extraction methods a lot of researches use Deep learning to extract the visual feature as done by Noda et al. [8] where CNN is used to extract the visual features to recognize phonemes the small part of sound [8] and Petridis [17] used LSTM on the extracted Deep Bottleneck Features (DBF) in combination with 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 of the conventional features is still a challenge for machine learning algorithms. Zhang et al. [18] 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 et al. [8] used HMM to recognize phonemes and Koller et al. [19] used it to recognize visemes the visual equivalent of the phonemes, also Goldschen et al. [20], Tao and Busso [1] used HMM in their Lipreading system. Support vector machines (SVMs) used by Zhao et al. [3] while Tamura et al. [21] used DBF to encode input images, Galatas et al. [22] used DCT, and Noda et al. [23] used CNN, all of them used these features with HMMs to classify spoken digits or isolated words. Salama et al. [14] introduced an AV-ASR 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 feature vector from both components then applied the HMM classifier. Deep learning achieved promising results in the classification stage in the speech recognition process. Mroueh et al. [24] employed feed-forward deep neural networks (DNNs) to implement phoneme classification using a non-public audio-visual dataset.

Deep learning has a lot of applications in real life, such as meteoric or brain diagnosing. In such cases, cognitive computing can help medical practitioners to diagnose patterns that they might not detectable, and they will extend the flexibility to diagnose the brain with efficiency [25]. Deep learning can be used in feature extraction and classification process, Petridis et al. [26] completed the previous work given in [27] 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 et al. [28] introduced a multimodal recurrent neural network (RNN) model to consider the sequential properties of audio and visual modalities for AV-ASR, 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 and Peleg [29] 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 [30] dataset. James et al. [31] 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 Graves et al. [32] which proved to improve the phoneme recognition performance.

Choosing large datasets that have many and variety of speakers is a challenging point in evaluating the performance of the AV-ASR 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 dataset to evaluate the performance of the proposed model but these researches either focus in specific speaker or perform phoneme recognition. Wand et al. [33], and Chung et al. [34] 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. Thanda and Venkatesan [35] proposed a training algorithm for an AV-ASR system using deep RNN which is evaluated on GRID corpus and provided a comparison of feature fusion and decision fusion. Recently, Shillingford et al. [36] performed labeling by using CNN, LSTM and Connectionist Temporal Classification (CTC) [37] which reports a strong speaker-independent performance on the constrained grammar and the 51 words vocabulary of GRID dataset. El Maghraby et al. [13] proposed an AV-ASR 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. 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 AV-ASR system [29, 38−40].

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

- Testing the proposed model with different size datasets to ensure its efficiency and using multi-speakers large dataset like GRID which suitable for learning stage of neural network.
- Integrating the visual speech feature with acoustic feature to design reliable AV-ASR system.
- Applying one of the most important DNN architecture BiLSTM in the classification process, and compares the obtained results with CNN and HMM.

Next section describes the proposed system which uses different classification techniques to test the enhancement that the visual features introduce to the recognition accuracy.

3.Proposed model

The proposed deep learning AV-ASR system is shown in *Figure 1*. There are three working stages: Data preparation stage, visual front-end and audio front-end. In the Data preparation stage, in case of GRID dataset, a synchronous audio file is extracted from the corresponding video file and segmenting them to the word boundary. Pre-processing and feature extraction operations are performed in audio and visual front-end separately. Then 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) or CNN to obtain the recognized words.

3.1Data 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.

3.1.1Extract alignment audio from video file

ffmpeg command is used 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.

3.1.2Word 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 [41]. 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 alignment file distributed with the dataset is used to get word level segmentations of the audio and video. For example, the given entry in alignment file “13500 20000 bin” give start and end time for the word “bin” in Nano second, so we segment the original audio and video files in this duration to get isolated word audio and video files using MATLAB [42] program. This segmentation caused the training dataset to consist of 6 words per sentence, for the 1000 sentences we have 6000 single words per speaker (each speaker has 1000 sentences). *Table 1* gives an example for the alignment of bbae9n.mpg file and the corresponding word boundary segmented video and audio files. The output of SFS program [43] explains the word boundary for bbae9n.wav file as shown in *Figure 2*.

Table 1 Segmented word boundary for audio and video files according to the corresponding alignment file for bbae9n.mpg

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

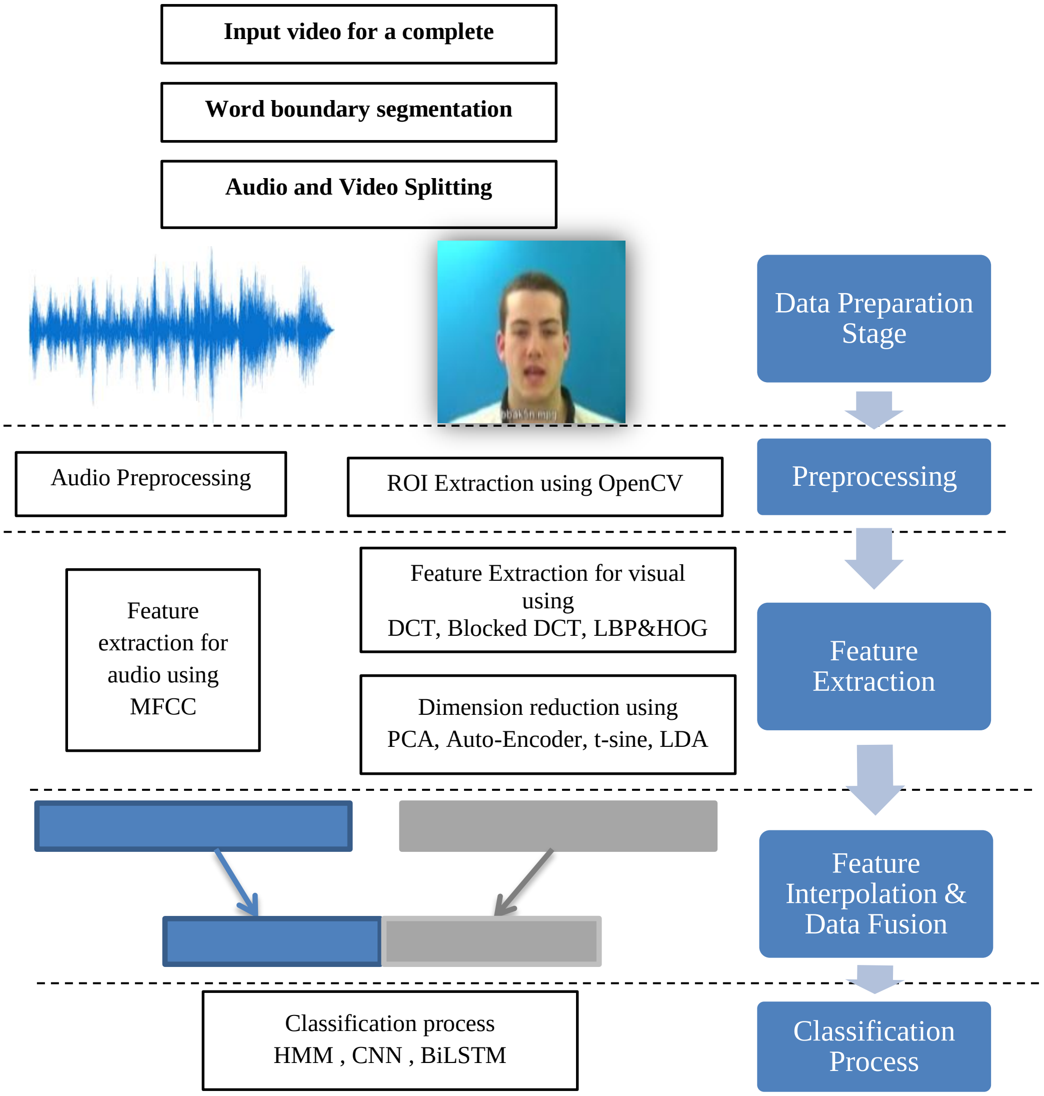
Input video for a complete
Word boundary segmentation
Audio and Video Splitting
bbak9n.mpg
Data Preparation Stage
Audio Preprocessing
ROI Extraction using OpenCV
Preprocessing
Feature extraction for audio using MFCC
Feature Extraction for visual using DCT, Blocked DCT, LBP&HOG
Dimension reduction using PCA, Auto-Encoder, t-sine, 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 [44]

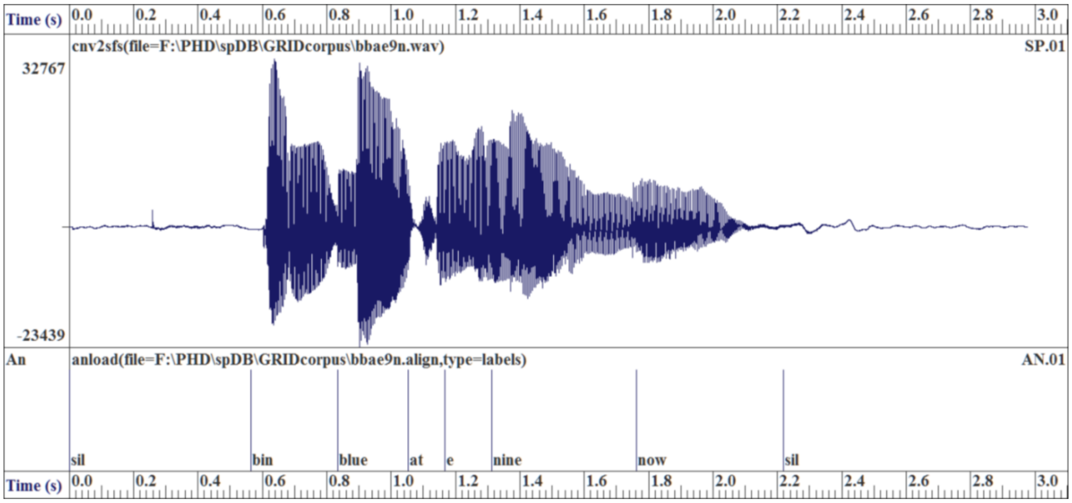

Figure 2 Word boundary according to alignment file, output from SFS [43] program

3.2Visual 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.

3.2.1Visual 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 [44] to detect and extract face and ROI for visual features is the speaker's mouth region by using Viola-Jones algorithm [45] 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 from speaker 12 of GRID dataset.

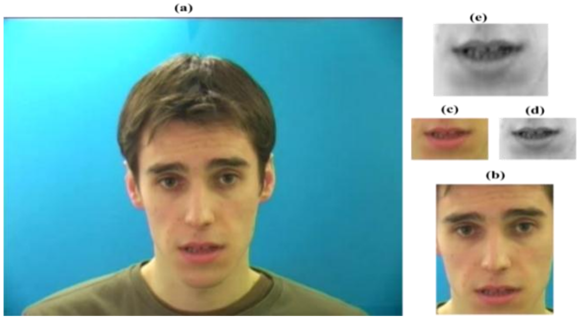

Figure 3 visual pre-processing steps for image from speaker 12 of GRID dataset. A) Original image, b) Detected face, c) Detected mouth, d) Mouth in grayscale, and e) Mouth after resizing 32x32

3.2.2Visual features extraction

There are two main visual feature extraction classes, appearance or pixel-based and shape or model-based. Model-based feature depends 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 all the required information. Appearance-based depends on all pixels in the mouth region and have valuable information to speech recognition [46]. Examples of appearance-based features methods are DCT, Discrete Wavelet Transform (DWT), and Linear Discriminate Analysis (LDA). In this research, we test using different visual features extraction methods like DCT, blocked DCT followed PCA and HOG with LBP features.

- **DCT** is used in this research due to its good performance in the previously discussed AV-ASR systems [47]. 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. Then, extract the final visual feature vector of 13 features from the upper left corner by using a zigzag scanning as shown in *Figure 4*. We try to increase or decrease the visual feature vector size around this value, but it gives the highest recognition rate.

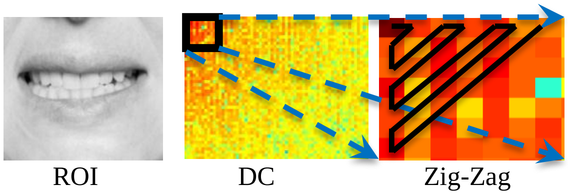

Figure 4 Zigzag scanning to extract feature vector from low frequency components of DCT matrix

- **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 criterion, such as 'low frequency'[13]. *Figure 5* explains the using strategy to extract the visual feature using blocked DCT with PCA Inspired by the cascade strategy by [48].

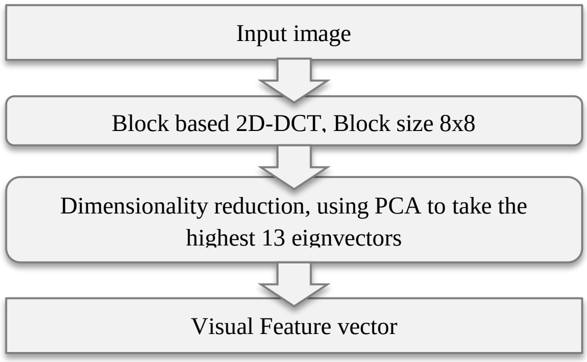
Input image
Block based 2D-DCT, Block size 8x8
Dimensionality reduction, using PCA to take the highest 13 eignvectors
Visual Feature vector

Figure 5 steps of visual feature extraction using blocked DCT

- **HOGs and LBPs:** HOG [49] and LBP have proven to be an effective descriptor for object recognition in general and face recognition in particular and can also be used in visual speech recognition. In [50] the authors successfully applied HOG descriptors to the problem of face recognition. In order 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 structure 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 [51].

3.3Audio front-end

Before extracting features from the audio file, it must be converted into small pieces called frames because speech signal is assumed to be stationary for a small period of time [14]. 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. After that, data is ready for the feature extraction step, Mel frequency Cepstral coefficient (MFCC) is the most effective audio feature extraction method [52] which simulate the variation of the human ear's important bandwidth with frequency. HMM Toolkit (HTK) [53] 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.

3.4Audio-visual features fusion

Fusion of audio and visual features can be divided into two categories: early integration (feature fusion), and late integration (decision fusion). In feature fusion, features from audio and visual are integrated to form a 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 is combined to take the final decision. In this paper early integration is performed by concatenating the audio and visual features. Before perform the concatenation between audio-visual features, we need to make sure that both must have the same feature extraction rates. The audio feature rate is greater than visual feature rate, which has the same rate as the video frame rate. As a result, 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 an audio vector) or 52 (13 from video vector+39 from an audio vector) is used for training and testing the classification stage.

3.5Classification

Previously HMM was the state of art classification technique for speech recognition systems for normal [13] and disorder people [14]. Although, HMM is easier to understand and implement, now deep learning gives much more accurate results for classification processes. Deep learning is preferred over HMM due to its robust self-learning mechanism and confirmed performance for speech recognition applications [15], although it would probably take more computational time but the results can be more reliable [16]. A lot of researches proved that the accuracy in certain image recognition and language processing problems is superior when using deep learning. In this paper, we compare the obtained results from the proposed model when using 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 CNN architecture is shown in *Figure 6*. The rectified linear unit (RELU) is used as 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 layer, the same configurations of the first convolution layer are used. Experiments were done on Adam optimizer and the loss function categorical cross-entropy.

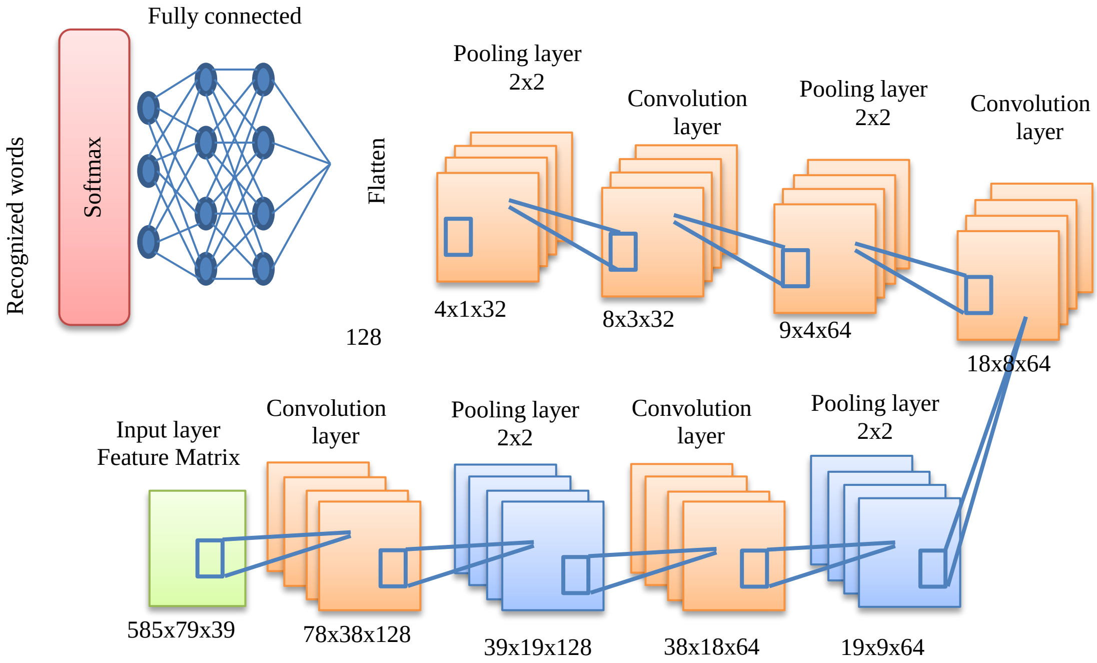

Figure 6 The proposed CNN architecture model

In the classification process, it is much more helpful to look at the frames of speech signal after it in addition to the previous frames, especially when it occurs close to the end of a word [32] to get the recognized word accurately. RNN is the most suitable for these cases. Feng et al. [28] explained the main structure of RNN, bidirectional RNNs, and LSTM which we use in building the classifier for our proposed AV-ASR 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 [32], in addition to, it's facing a challenging problem known as the gradient vanishing and exploding problem [54]. The solution for these problems is to use bidirectional LSTM [55, 56]. LSTM model firstly proposed by Hochreiter and Schmidhuber [57], it confirms that the gradients can pass through many time steps and overcomes both vanishing and exploding insufficiencies of the gradients [28]. Bidirectional recurrent neural (BiRNN) networks introduced by Schuster and Paliwal [55], it consists of two RNNs forward and backward ones. Because of its good results, BiRNN has been used in speech recognition [56, 58], and handwriting recognition [59, 60] systems. In combination with the benefits of BiRNN and the enhancement that LSTM introduces [56] BiLSTM can be the optimal solution for the classification process in speech recognition. In our proposed model, as shown in *Figure 7* the extracted features are used as the input layer, then two copies of the hidden layer of LSTM are created, one fit in the input sequences as-is and one with 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 previously system, we decide to use BiLSTM in compared to CNN and HMM for the classification process in our model. All the used hyper-parameters of the BiLSTM model are shown in *Table 2*.

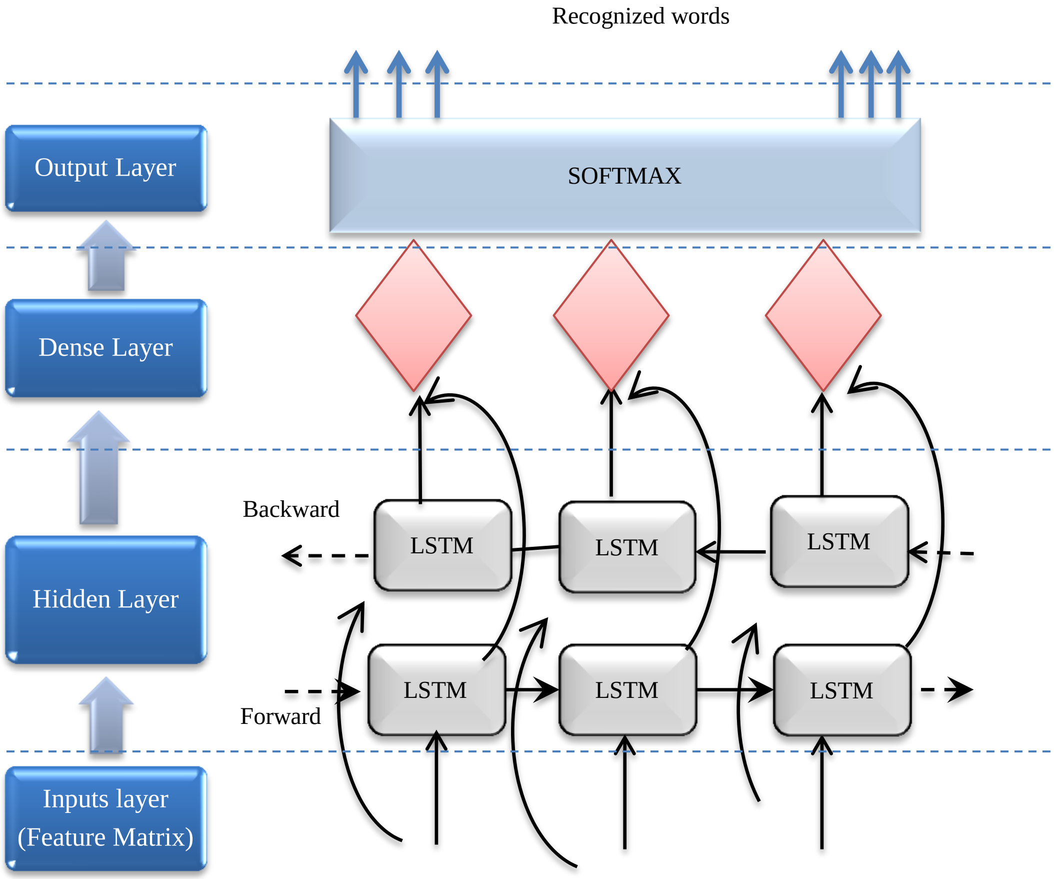

Figure 7 BiLSTM proposed classifier structure

Table 2 Hyper-parameters for the BiLSTM classifier

Hyper-parameters	Values
No. of memory units	100
BiLSTM layer activation	SIGMOID
Dense layer activation	SOFTMAX
Optimizer	Adam
Loss	Categorical Cross-Entropy

In case of the HMM classifier HMM toolkit (HTK) [53] is used for organizing, training and testing the HMM model. A total of 51 HMM models, one for each word, are trained for Grid dataset and 26 HMM models for AVLetters dataset. 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.

4.Procedure and system model

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

4.1Datasets

GRID audio-visual corpus is used in the training and the testing stages. It is a collection of audio and video recordings for 34 speakers (18 males, 16 females) ages ranged from 18 to 49 years each saying 1000 sentences [30]. The total length of the recordings is 28 hours; with total number of words in the vocabulary 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 words representing command, 4 colors, 4 prepositions, 26 letters, 10 digits, and 4 adverbs as listed in *Table 3*. The video was recorded as a sequence of images with a frame rate of 25fbs.

Table 3 Sentence structure for the GRID corpus [30]

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

Also, to check the performance of the proposed model for different size dataset, our model is tested in a small benchmark dataset like AVletters [2], where ten different speakers repeat the isolated letters A-Z three times, a total of 10x3x26=780 video files. *Figure 8* shows the grammar for the GRID corpus for isolated word. Also *Figure 9* illustrates the grammar for AVletter dataset.

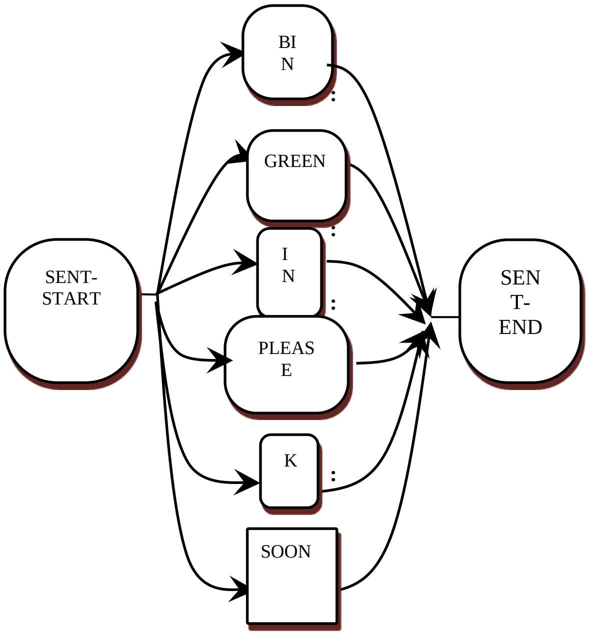

Figure 8 Grammar for word recognition, GRID dataset

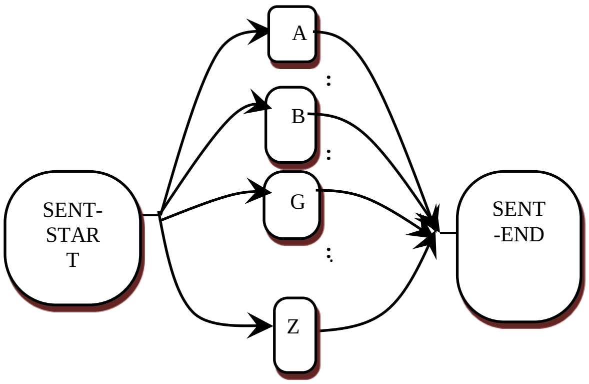

Figure 9 Grammar for word recognition, AVletter dataset

4.2Performance analysis

In case of HMM classifier, the system performance is analyzed using Hresults tool of HTK [53], it used to calculate the recognition accuracy of the speech system, evaluated as:

$$\%\ \text{Accuracy} = 100 \times \frac{(N - D - S - I)}{N} \qquad (1)$$

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 the evaluation process of the proposed model and compares our methods to the previous state of the art. The results for GRID and AVletters datasets 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.

5.1Avletters results

We evaluate the proposed model firstly on the AVletters benchmark dataset. MFCC is used to extract the audio features with feature vector size of 26, and DCT is used to extract the visual features with feature vector size of 13. *Table 4* shows the results of using two major DNN architectures (CNN and BiLSTM) in the classification process. The recognition accuracies of using HMM are shown in *Table 5* with different Gaussian mixtures, the gray cells in theses tables are an indication of the highest recognition values. *Figure 10* compares the results obtained by using different classifier methods and feature types to identify the best model which gives the highest recognition accuracy.

Table 4 % Accuracy result of CNN and BiLSTM for AVletters using video only, audio-only, and audio-visual features

DNN architecture **Feature type**	**BiLSTM**	**CNN**
A	89.23%	63.5897%
V	85.13%	**90.256%**
AV	**93.33%**	84.1026 %

Table 5 % Accuracy result of HMM for AVletters using (V) video only, (A) audio-only, and (AV) audio-visual features with different Gaussian

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

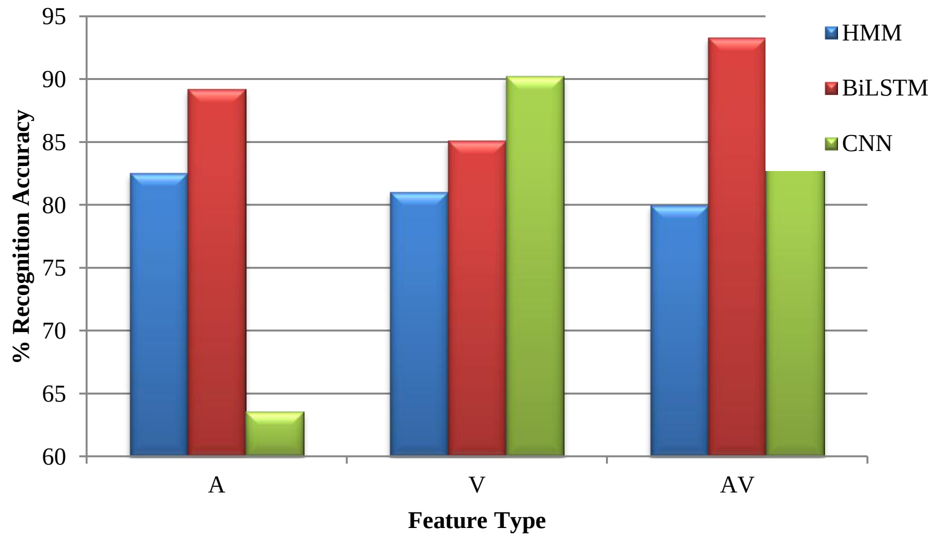

Figure 10 %Recognition accuracy results for A, V, and AV features and different classifiers HMM, CNN and BiLSTM

Based on the obtained results, we found that:

- Using BiLSTM with integrated audio-visual feature gives enhancement over audio-only by 8.33% and decreases the loss value by 45.57%.
- Using CNN with integrated audio-visual features gives enhancement over audio-only by 32.3% and decreases the loss by 24.8%.
- In case of HMM, Increasing the Gaussian mixtures in HMM enhances the recognition accuracy until reach 8 mixture after that it decreases. The best result is 82.56% with 4 mixture in audio only, 77.95% for video-only, and 80% for audio-video.

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 it obtained in [28] which gives an accuracy of 87.7% for audio-visual using the same dataset as shown in *Table 6*. DCT is used to extract the main important features from the input image then selecting the main important features using zigzag scanning (minimized numbers of features) then feeding these features to BiLSTM classifier to perform video-only speech recognition. The proposed system has advantage over system introduced in [28] where using DCT has the ability to efficiently represent the mouth region using a fixed number of coefficients, which leads to fast matching algorithms. Also using zigzag scanning minimizes the vector size without several processing.

Table 6 Comparison between best recognition accuracy obtained in our model and the previously obtained results in [28]

	Feng. W [28]	**Ours**
Audio-only	75.6%	**86.15%**
Visual-only	64.4%	**85.13%**
Audio-visual	87.7%	**93.33%**

The confusion matrix for the BiLSTM AV-ASR model is shown in *Figure 11* for AVletter database, it explains the relation between true and recognized word where the Greyscale level indicates the density of matching between them.

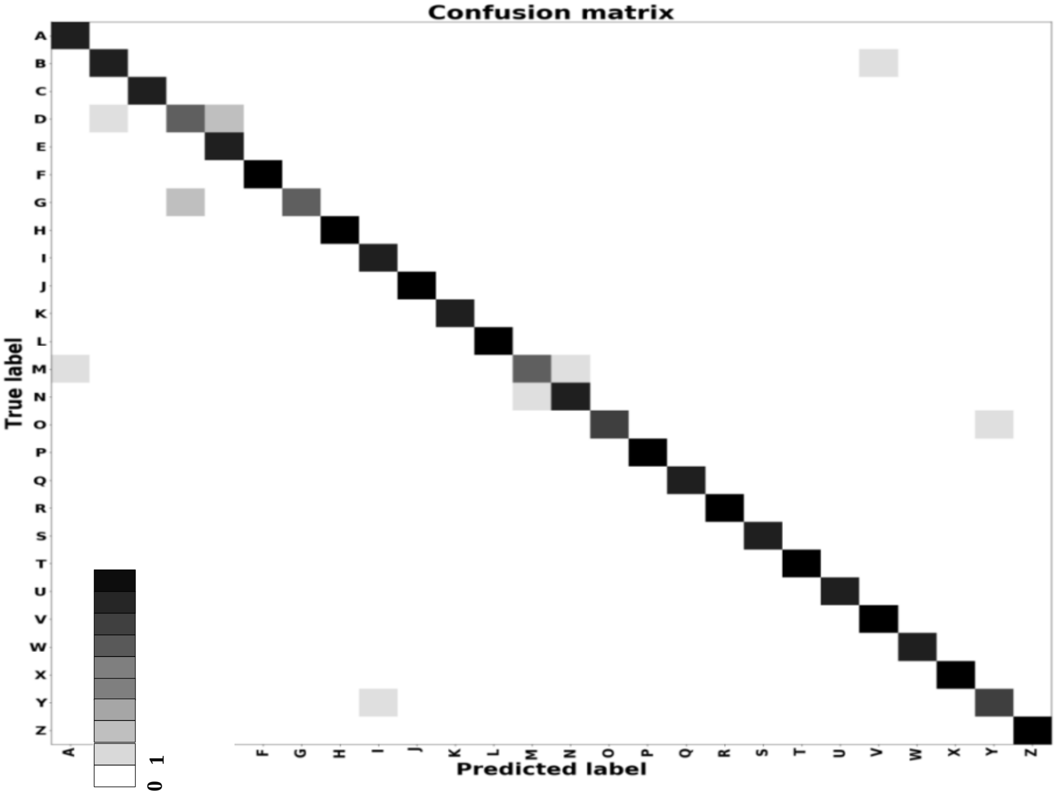

Figure 11 AVletters confusion matrix with audio-visual feature and LSTM classifier

5.2GRID 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.

The experiments performed in GRID dataset is divided into three stages audio-only with different feature vector size, visual-only with different feature extraction and dimension reduction techniques, and AV-ASR to ensure the improvement of the addition of both feature to the recognition accuracy. Firstly, the experiment performed for speaker 4 from the GRID database to compare the performance of the model to results obtained in [29].

5.2.1Visual-only speech recognition

Table 7 shows a comparison between the performance of the HMMv, CNNv and BiLSTMv VSR models after utilizing DCT or Blocked DCT or HOG+LBP features in the visual front-ends. It 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. The utilization of DCT in BiLSTMv model termed as DCT- BiLSTMv, significantly outperforms the traditional DCT-HMMv model by a very large percentage difference 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.

Table 7 % Accuracy results of the HMMa, CNNa, and BiLSTMa models with video-only (Vd) DCT or (Vbd) blocked DCT or (Vhlbp) HOG+LBP as input features

	HMMv	**BiLSTMv**	**CNNv**
VD	52.47 mix16	**78.87**	75.6
VBD	23.33 mix64	**46.8**	44.26667
VHLBP	27.07 mix64	**42.4**	40.53333

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 [29], we initially performed our experiments on speaker four (S4, female) as done there.

5.2.2Audio-only speech recognition

Table 8 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 performance of the ASR 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 til reach mix8 after that decrease, the improvement of A39-BiLSTMa become more obvious when adding the noise signal. The results show that the deep A39-BiLSTMa ASR model outperforms the other ASR models.

5.2.3Audio-visual speech recognition

Figures 12 and *13* shows a comparison between the performance of the 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. It demonstrates that the performance of the AV-ASR 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 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.

Table 8 % Accuracy results of the HMMa, CNNa and BiLSTMa models for audio-only A13 or A39 as input features, with different SNR

Classification method	HMMa		BiLSTMa		CNNa	
Feat. type / SNR	A13	A39	A13	A39	A13	A39
CLEAN	89.51 Mix8	95.93	**98.2**	97.47	95.53	95.93
N20	87.94 Mix8	95.4	98.13	**97.36**	95.33	95.4
N15	87.26 Mix8	95.27	97.73	**97.2**	95.23	95.27
N10	87.6 Mix8	95.07	95.47	**97.23**	95.13	95.07
N5	86.78 Mix8	93.53	95	**97**	95.04	93.53
N0	86.24 Mix16	92.27	94.13	**95.73**	93.53	92.27

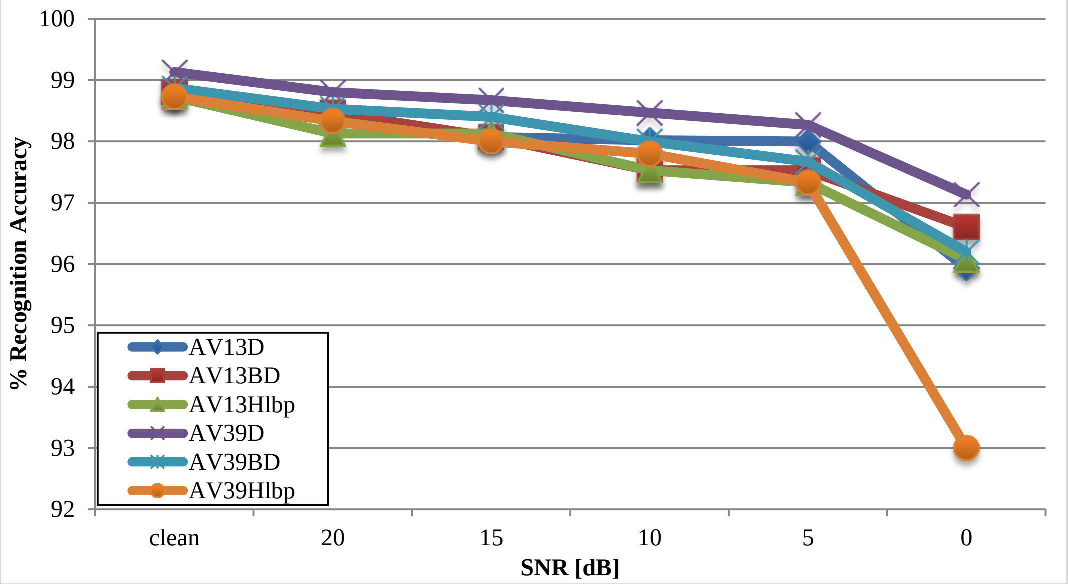

Figure 12 Comparison between the performance of different combinations for audio and visual features with the BiLSTMav classifier. AV13D, AV13BD, and AV13HLbp stands for early integrated MFCC_0 with DCT, blocked DCT and HOG+LBP features respectively, while AV39D, AV39BD, and AV39HLbp stand for early integrated MFCC_D_A_0 with DCT, blocked DCT and HOG+LBP features respectively

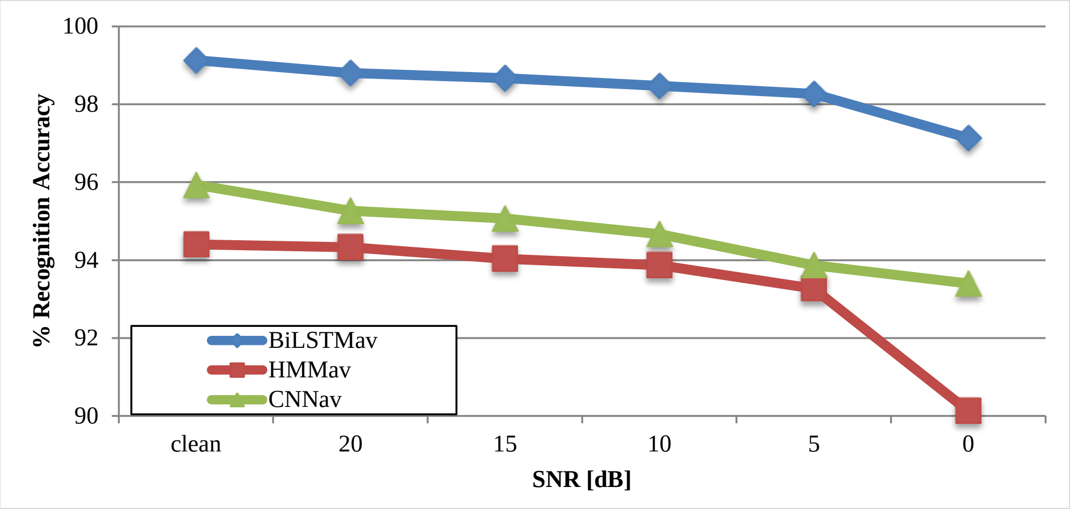

Figure 13 Performance comparison of deep BiLSTMav, HMMav, and CNNav models using MCC_D_A_0 as audio features and DCT as visual features at different SNR levels

5.2.4Result for different dimension reduction methods

To reduce the visual feature vector size without losing the main feature. There a lot of dimension reduction techniques can be used and test their effect on the recognition accuracy. The result of using different dimension reduction techniques like PCA, auto-encoder, LDA, t-SNE is shown in *Table 9.*

Matlab Toolbox for Dimensionality Reduction [39] 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 9 % Recognition accuracy for different dimension reduction techniques when using HMM classification technique

		mono	**tri**	**mix2**	**mix4**	**mix8**	**mix16**	**mix32**	**mix64**	**mix128**
AV_DCT_tsine	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 up to 24% and for audio-only up to 18.8% for speaker 4 when using the BiLSTM classifier which is better than the recognition accuracy achieved by Ephrat [29] for the same speaker.

Our proposed model gives recognition accuracy for speaker 4 is 99.13% with BiLSTM classifier which is better than 79.9% obtained by Ephrat [29] for the same dataset and same speaker, because our model depends on BiLSTM instead of CNN which used in [29]. *Table 10* gives the comparison between the performances of our model and Ephrat model for speaker 4 using audio-only and audio-visual signal.

Table 10 Comparison between the best recognition accuracy obtained by our model and the previously obtained results in [29]

	Ephrat [29]	**Ours**
Audio-only	82.6%	**98.2%**
Audio-visual	79.9%	**99.13%**

We also perform experiments on speakers 6, 11, and 12 (two females and two males) to ensure the robustness of our model.

From the obtained result illustrated in *Table 11*, 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 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 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 signal and when using either CNN or BiLSTM classification methods.

Table 11 % Accuracy results for speakers 6, 11, and 12 for GRID dataset

			A13	A39	V13	AV13	AV39
Clean	**BiLSTM**	S6	98.90%	99.10%	93.40%	99.60%	99.70%
		S11	96.30%	96.70%	90.30%	97.80%	98.20%
		S12	99.30%	99%	93.90%	99.70%	99.80%
	CNN	S6	97%	97%	97.40%	98.50%	98.70%
		S11	91.20%	88.70%	96.50%	95.80%	96.90%
		S12	96.20%	95.90%	97.10%	98.70%	98.90%
	HMM	S6	94.54 mix16	94.47 mix16	94.34 mix16	97.71 mix16	88.22 mix16
		S11	86.7% mix8	87.9% mix8	80.3% mix8	88.7% mix8	89.9% mix8
		S12	95.2% mix8	95.9% mix8	84.1% mix8	91.5% mix8	97.3% mix8
Noise 5dB	**BiLSTM**	S6	97.60%	98.50%	93.40%	99.50%	99.50%
		S11	94.10%	93.90%	90.30%	96.10%	96.90%
		S12	98.50%	98.60%	93.90%	99.50%	99.60%
	CNN	S6	95.50%	90.80%	97.40%	97.50%	96.40%
		S11	88%	85.70%	96.50%	95.40%	96.50%
		S12	94.90%	93.10%	97.10%	98.80%	98.80%
	HMM	S6	91.12 mix16	92.4% mix16	94.34 mix16	85% mix16	96% mix16
		S11	74.4% mix8	79.7% mix8	80.3% mix8	80.9% mix8	79.1% mix8
		S12	91.7% mix4	93.3% mix4	84.1% mix8	88.5% mix4	94.9% mix4

Figure 14 gives 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 12. 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 15 introduces the confusion matrix for speaker 12 of GRID database, using BiLSTM AV-ASR model 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".

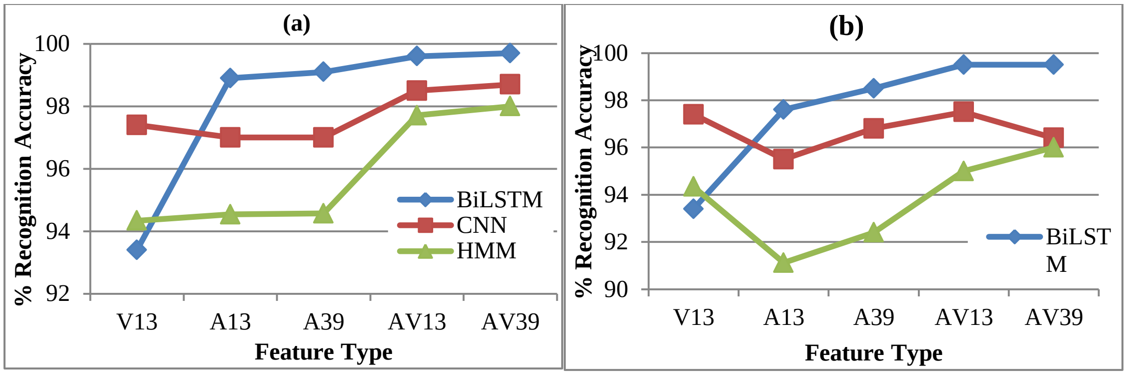

Figure 14 Result of using different classification techniques with audio-only feature MFCC_0 (A13), MFCC_D_A_0 (A39), visual-only (V13) and audio-visual feature MFCC_0+V13 (AV13), MFCC_D_A_0+V13 (AV13) in (a) clean, (b) noisy system

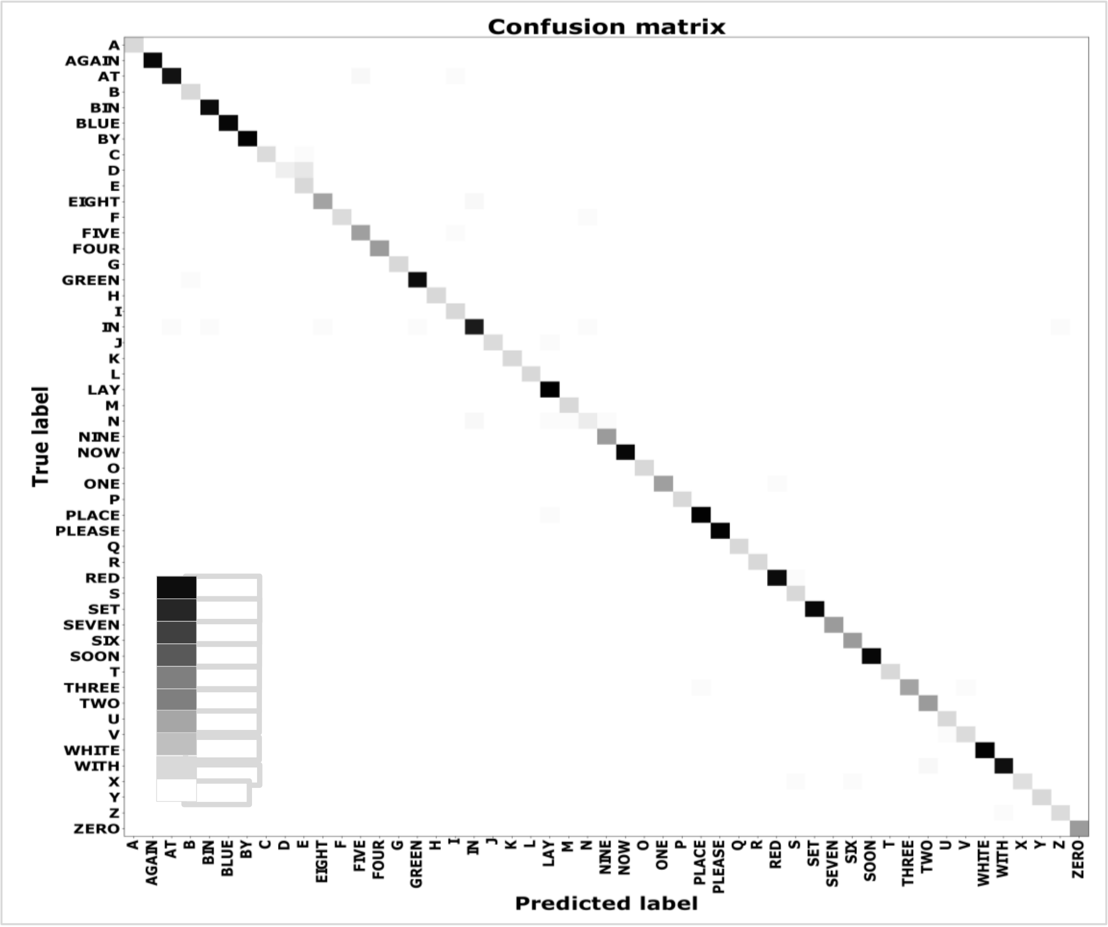

Figure 15 Isolated word recognition confusion matrix for Speaker 12 of the GRID dataset AV39-BiLSTM model

6.Current and future developments

In this research, an isolated word speech recognition system is demonstrated from the early integration of multimodal features extracted by using MFCC and DCT for audio and visual signal respectively. The classification process is performed by using BiLSTM, CNN, and HMM to select the most reliable classification techniques.

However, our current results from the GRID database are obtained by preparing speaker-dependent features for CNN, BiLSTM, and HMM model. Although

DNN (BiLSTM and CNN) framework is scalable, it requires sufficient training dataset to reduce overfitting [61] where the number and variation of training samples are very important issues for improving the generalization ability of a DNN. Therefore, in future work, we need to check the possibility of constructing an AV-ASR system applicable to the real-world environment by training and evaluating our current model with a various audio-visual speech dataset that has large variations in images where dynamic changes such as reverberation, illumination, and facial orientation, occur [28].

In future, it might be interesting to formulate an AV-ASR model based on the using of late integration (decision fusion) instead of early integration that used in this research can give better recognition accuracy which enables the different modality to use the classification technique that suitable for the modality type of the signal. In the classification process, it may be useful to use an integration of DNN and HMM approach because of the recognition capability of DNNs and the simplicity of the proposed late integration approach.

7.Data availability

The datasets analyzed during the current study are introduced in [2] and [30]. These datasets are available online for research topic 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/

8.Conclusion

This work introduces an AV-ASR model based on the deep learning approach as a solution for designing reliable and noise robust speech recognition system. The proposed model comprised into two main stages: 1) extracting effective feature from audio and visual signal separately where MFCC is used for the audio signal with different vector sizes either 13 or 39. According to the visual feature, comparison between different feature extraction methods is performed to select the most effective visual features from the lip region like DCT or Blocked DCT or HOG+LBP. Then different dimension reduction techniques are applied to select the most important visual features from the previously extracted feature using either PCA or t-sine or auto-encoder or LDA. 2) Applying different classification methods like BiLSTM, CNN and HMM to obtain the final recognition decision. The proposed model is evaluated on two multi-speakers' audio-visual datasets (AVletter and GRID) with speaker-dependent and independent experiments to ensure the performance enhancement of the proposed model. The experimental results showed that: firstly, according to the visual signal, extracting the visual features using DCT with zigzag scanning is the optimal visual feature extraction and dimension reduction technique when compared to other used methods because DCT can compactly represent the mouth region using fixed number of coefficients, which lead to fast matching algorithms. Secondly, applying different classification techniques like BiLSTM, CNN and traditional method HMM proved that BiLSTM is the most suitable classifier used to obtain reliable and noise-robust speech recognition system. Finally, combining DCT visual feature and MFCC audio feature in addition to the using 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. Comparing the proposed model to previously obtain results which using the same datasets, we found that our model gives higher recognition accuracy.

Acknowledgment

None.

Conflicts of interest

The authors have no conflicts of interest to declare.

References

[1] Tao F, Busso C. Lipreading approach for isolated digits recognition under whisper and neutral speech. In fifteenth annual conference of the international speech communication association 2014 (pp.1154-8).

[2] Matthews I, Cootes TF, Bangham JA, Cox S, Harvey R. Extraction of visual features for lipreading. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2002; 24(2):198-213.

[3] Zhao G, Barnard M, Pietikainen M. Lipreading with local spatiotemporal descriptors. IEEE Transactions on Multimedia. 2009; 11(7):1254-65.

[4] Petajan ED. Automatic lipreading to enhance speech recognition (Speech Reading).1985.

[5] Neti C, Potamianos G, Luettin J, Matthews I, Glotin H, Vergyri D, et al. Audio-visual speech recognition final workshop report. Center for language and speech processing, Johns Hopkins University, Baltimore 2000.

[6] Potamianos G, Graf HP, Cosatto E. An image transform approach for HMM based automatic lipreading. In proceedings international conference on image processing, ICIP98 (Cat. No. 98CB36269) 1998 (pp. 173-7). IEEE.

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

[8] Noda K, Yamaguchi Y, Nakadai K, Okuno HG, Ogata T. Lipreading using convolutional neural network. In fifteenth annual conference of the international speech communication association 2014 (pp.1149-53).

[9] Chowdhary CL. 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 2018 (pp. 218-31). IGI Global.

[10] Chan MT. HMM-based audio-visual speech recognition integrating geometric-and appearance-based visual features. In fourth workshop on multimedia signal processing (Cat. No. 01TH8564) 2001 (pp. 9-14). IEEE.

[11] McGurk H, MacDonald J. Hearing lips and seeing voices. Nature. 1976; 264:746-8.

[12] Potamianos G, Neti C, Gravier G, Garg A, Senior AW. Recent advances in the automatic recognition of audiovisual speech. Proceedings of the IEEE. 2003; 91(9):1306-26.

[13] El Maghraby EE, Gody AM, Farouk MH. Enhancing quality and accuracy of speech recognition system by using multimodal audio-visual speech signal. In international computer engineering conference 2016 (pp. 219-29). IEEE.

[14] Salama ES, El-Khoribi RA, Shoman ME. Audio-visual speech recognition for people with speech disorders. International Journal of Computer Applications. 2014; 96(2):51-6.

[15] Schmidhuber J. Deep learning in neural networks: an overview. Neural Networks. 2015; 61:85-117.

[16] Hinton G, Deng L, Yu D, Dahl GE, Mohamed AR, Jaitly N, et al. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Processing Magazine. 2012; 29(6):82-97.

[17] Petridis S, Pantic M. Deep complementary bottleneck features for visual speech recognition. In international conference on acoustics, speech and signal processing 2016 (pp. 2304-8). IEEE.

[18] Zhang F, Li W, Zhang Y, Feng Z. Data driven feature selection for machine learning algorithms in computer vision. IEEE Internet of Things Journal. 2018; 5(6):4262-72.

[19] Koller O, Ney H, Bowden R. Deep learning of mouth shapes for sign language. In proceedings of the international conference on computer vision workshops 2015 (pp. 477-83).

[20] Goldschen AJ, Garcia ON, Petajan ED. Continuous automatic speech recognition by lipreading. In motion-based recognition 1997 (pp. 321-43). Springer, Dordrecht.

[21] Tamura S, Ninomiya H, Kitaoka N, Osuga S, Iribe Y, Takeda K, et al. Audio-visual speech recognition using deep bottleneck features and high-performance lipreading. In asia-pacific signal and information processing association annual summit and conference 2015 (pp. 575-82). IEEE.

[22] Galatas G, Potamianos G, Makedon F. Audio-visual speech recognition incorporating facial depth information captured by the Kinect. In proceedings of the European signal processing conference 2012 (pp. 2714-7). IEEE.

[23] Noda K, Yamaguchi Y, Nakadai K, Okuno HG, Ogata T. Audio-visual speech recognition using deep learning. Applied Intelligence. 2015; 42: 722-37.

[24] Mroueh Y, Marcheret E, Goel V. Deep multimodal learning for audio-visual speech recognition. In international conference on acoustics, speech and signal processing 2015 (pp. 2130-4). IEEE.

[25] Chowdhary CL, Darwish A, Hassanien AE. Cognitive deep learning: future direction in intelligent retrieval. In handbook of research on deep learning innovations and trends 2019 (pp. 220-31). IGI Global.

[26] Petridis S, Stafylakis T, Ma P, Cai F, Tzimiropoulos G, Pantic M. End-to-end audiovisual speech recognition. In international conference on acoustics, speech and signal processing 2018 (pp. 6548-52). IEEE.

[27] Stafylakis T, Tzimiropoulos G. Combining residual networks with LSTMs for lipreading. arXiv preprint arXiv:1703.04105. 2017.

[28] Feng W, Guan N, Li Y, Zhang X, Luo Z. Audio visual speech recognition with multimodal recurrent neural networks. In international joint conference on neural networks 2017 (pp. 681-8). IEEE.

[29] Ephrat A, Peleg S. Vid2speech: speech reconstruction from silent video. In IEEE international conference on acoustics, speech and signal processing 2017 (pp. 5095-9). IEEE.

[30] Cooke M, Barker J, Cunningham S, Shao X. An audio-visual corpus for speech perception and automatic speech recognition. The Journal of the Acoustical Society of America. 2006; 120(5):2421-4.

[31] James PE, Mun HK, Vaithilingam CA. A hybrid spoken language processing system for smart device troubleshooting. Electronics. 2019; 8(6):1-16.

[32] Graves A, Fernández S, Schmidhuber J. Bidirectional LSTM networks for improved phoneme classification and recognition. In international conference on artificial neural networks 2005 (pp. 799-804). Springer, Berlin, Heidelberg.

[33] Wand M, Koutník J, Schmidhuber J. Lipreading with long short-term memory. In international conference on acoustics, speech and signal processing 2016 (pp. 6115-19). IEEE.

[34] Chung JS, Senior A, Vinyals O, Zisserman A. Lip reading sentences in the wild. In conference on computer vision and pattern recognition 2017 (pp. 3444-53). IEEE.

[35] Thanda A, Venkatesan SM. Audio visual speech recognition using deep recurrent neural networks. In IAPR workshop on multimodal pattern recognition of social signals in human-computer interaction 2016 (pp. 98-109). Springer, Cham.

[36] Shillingford B, Whiteson S, Assael ND. Lipnet: sentence-level lipreading. In GPU technology conference 2016.

[37] Graves A, Fernández S, Gomez F, Schmidhuber J. Connectionist temporal classification: labelling unregimented sequence data with recurrent neural networks. In proceedings of the international conference on machine learning 2006 (pp. 369-76).

[38] Barker J, Vincent E, Ma N, Christensen H, Green P. The PASCAL CHiME speech separation and recognition challenge. Computer Speech & Language. 2013; 27(3):621-33.

[39] Gan T, Menzel W, Yang S. An audio-visual speech recognition framework based on articulatory features. Auditory-Visual Speech Processing 2007.

[40] Cornu TL, Milner B. Reconstructing intelligible audio speech from visual speech features. In sixteenth annual conference of the international speech communication association 2015 (pp. 3355-9).

[41] Bear HL, Harvey R. Decoding visemes: improving machine lip-reading. In international conference on acoustics, speech and signal processing 2016 (pp. 2009-13). IEEE.

[42] http://www.mathworks.com. Accessed 20 October 2019.

[43] https://www.phon.ucl.ac.uk/. Accessed 20 October 2019.

[44] http://www.opencv.org/ . Accessed 20 October 2019.

[45] Jensen OH. Implementing the viola-jones face detection algorithm. Masters thesis, Technical University of Denmark, DTU, DK-2800 Kgs. Lyngby, Denmark. 2008.

[46] Potamianos G, Scanlon P. Exploiting lower face symmetry in appearance-based automatic speechreading. In AVSP 2005 (pp. 79-84).

[47] Estellers V, Thiran JP. Multi-pose lipreading and audio-visual speech recognition. EURASIP Journal on Advances in Signal Processing. 2012.

[48] Nefian AV, Liang L, Pi X, Liu X, Murphy K. Dynamic bayesian networks for audio-visual speech recognition. EURASIP Journal on Advances in Signal Processing. 2002.

[49] Lowe DG. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision. 2004; 60(2):91-110.

[50] Albiol A, Monzo D, Martin A, Sastre J, Albiol A. Face recognition using HOG–EBGM. Pattern Recognition Letters. 2008; 29(10):1537-43.

[51] Ghorbani M, Targhi AT, Dehshibi MM. HOG and LBP: towards a robust face recognition system. In tenth international conference on digital information management 2015 (pp. 138-41). IEEE.

[52] Tiwari V. MFCC and its applications in speaker recognition. International Journal on Emerging Technologies. 2010; 1(1):19-22.

[53] Young S, Evermann G, Gales M, Hain T, Kershaw D, Liu X, et al. The HTK book. Cambridge University, Engineering Department. 2006.

[54] Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks. 1994; 5(2):157-66.

[55] Schuster M, Paliwal KK. Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing. 1997; 45(11):2673-81.

[56] Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks. 2005; 18(5-6):602-10.

[57] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation. 1997; 9(8):1735-80.

[58] Graves A, Mohamed AR, Hinton G. Speech recognition with deep recurrent neural networks. In international conference on acoustics, speech and signal processing 2013 (pp. 6645-9). IEEE.

[59] Graves A, Liwicki M, Bunke H, Schmidhuber J, Fernández S. Unconstrained on-line handwriting recognition with recurrent neural networks. In advances in neural information processing systems 2008 (pp. 577-84).

[60] Graves A, Liwicki M, Fernández S, Bertolami R, Bunke H, Schmidhuber J. A novel connectionist system for unconstrained handwriting recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2008; 31(5):855-68.

[61] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In advances in neural information processing systems 2012 (pp. 1097-105).

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 in Information Systems Department, Faculty of Computers and Information, Fayoum University. Her research interest is in Signal Processing and Computer Networks.
Email: eem00@fayoum.edu.eg

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.
Email: amg00@fayoum.edu.eg

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 Department, Faculty of Engineering, Cairo University from 2007-Till Now. He is author and co-author of about 40 papers in National and International conference proceedings and journals.
Email: mhesham@eng.cu.edu.eg