Computers, Materials & Continua
DOI:10.32604/cmc.2022.019450
Article

Tech Science Press

Deep Learning-Based Approach for Arabic Visual Speech Recognition

Nadia H. Alsulami[1,*], Amani T. Jamal[1] and Lamiaa A. Elrefaei[2]

[1]Computer Science Department, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, 21589, Saudi Arabia
[2]Electrical Engineering Department, Faculty of Engineering at Shoubra, Benha University, Cairo, 11629, Egypt
[*]Corresponding Author: Nadia H. Alsulami. Email: nadiaalsulami@yahoo.com
Received: 14 April 2021; Accepted: 03 June 2021

Abstract: Lip-reading technologies are rapidly progressing following the breakthrough of deep learning. It plays a vital role in its many applications, such as: human-machine communication practices or security applications. In this paper, we propose to develop an effective lip-reading recognition model for Arabic visual speech recognition by implementing deep learning algorithms. The Arabic visual datasets that have been collected contains 2400 records of Arabic digits and 960 records of Arabic phrases from 24 native speakers. The primary purpose is to provide a high-performance model in terms of enhancing the preprocessing phase. Firstly, we extract keyframes from our dataset. Secondly, we produce a Concatenated Frame Images (CFIs) that represent the utterance sequence in one single image. Finally, the VGG-19 is employed for visual features extraction in our proposed model. We have examined different keyframes: 10, 15, and 20 for comparing two types of approaches in the proposed model: (1) the VGG-19 base model and (2) VGG-19 base model with batch normalization. The results show that the second approach achieves greater accuracy: 94% for digit recognition, 97% for phrase recognition, and 93% for digits and phrases recognition in the test dataset. Therefore, our proposed model is superior to models based on CFIs input.

Keywords: Convolutional neural network; deep learning; lip reading; transfer learning; visual speech recognition

1 Introduction

Lip-reading recognition system which is also known as Visual Speech Recognition (VSR), plays an essential role in human language communication and visual knowledge. It refers to the ability to learn or recognize visual speech without the need to hear the audio, and it works only with visual data (such as movements of the lips and face). Lip-reading technology is an appealing area of study for researchers because, by recognizing visual information without the need of audio, it introduces a new tool in visual speech recognition for situations in which audio is not available or must be secured professionally. Recognizing spoken words from the speaker's lip movement is called visual lip-reading, and it is an efficacious communication form in many situations. People with hearing disabilities, for example, can be served by this useful hearing aid [1]. The automatic

recognition of visual speech greatly improves human communication by helping handicapped persons via interaction with the machine through the visual information from the speaker's lip, enhancing human speech perception by extracting the visual features. This technology includes the following necessary processes: face detection, lip localization, feature extraction, training the classifier, and recognition of the word [2], as demonstrated in Fig. 1.

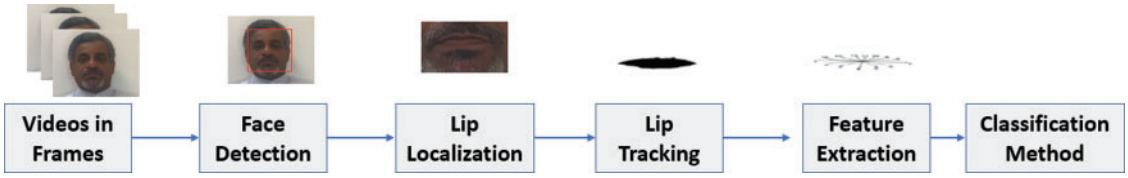

Figure 1: Visual speech recognition main processes

Lip visual feature extraction is considered the most critical phase in preparing the extracted visual data for the automatic lip-reading model. The existing feature extraction approaches can be classified into two main categories: handcrafted features-based models and deep learning features-based models. In the handcrafted features-based models the geometrical, model-based, appearance-based, image-transformed, and motion-based features are applied to manually extract features as demonstrated in [3]. This approach uses a set of rules and algorithms, including the shape of the lip, contour, mouth height or width, etc. As reviewed by [4], dimensionality reduction techniques have been utilized in handcrafted approaches as in the transformed approaches where images are transformed into a space of feature such as Discrete Cosine Transform (DCT), and the Principal Component Analysis (PCA). Recognition of the spoken words is the next step in which the extracted features have usually been supplied to the classifiers, such as the Hidden Markov Model (HMM) [5]. Recently, deep learning had meaningful artificial progress in the computer vision field. Accordingly, researchers have been encouraged to alter their research direction to focus on lip-reading approaches that are forwarded to deep learning models rather than the traditional manual feature extraction [6]. Recently, researchers have been focused on Convolutional Neural Networks (CNN) to concentrate on: region of interest, classification, and the purpose of image detection. Due to the continuous changes in lip-reading visual data, some researchers intend to apply the Long Short-Term Memory (LSTM) network as will be discussed in the reviewed studies. Indeed, the handcrafted process requires extra knowledge, tedious work by hand, and input data knowledge, which needs more time. In contrast, deep learning-based models avoid these processes, and their basic property is automatic learning for a series of nonlinear features that enhance the performance of many tasks in computer vision [7].

Nowadays, researchers intended to replace any spoken word with CFI, as input representation in the CNN model, where each video converts into one single image. Some researchers employed the concept of stretching the sequence with the nearest image to a specific length to normalize the speaking speed, and others have produced the CFI with a fixed number of frames [8].

The importance of this research is due to the limited studies for developing a lip-reading technology in the Arab region particularly in the deep learning area. Therefore, this paper aims to develop a lip-reading model for the Arabic word recognition, to classify the lip images extracted from the video database by applying a high-performance features extraction method, and by utilizing CNN to achieve a high accuracy recognition of Arabic words.

The key contributions of this research are:

1) Building an Arabic visual dataset to evaluate and test the proposed model, the videos were collected from different native speakers, taking into consideration all the suitable environments.

2) Applying the concept of the concatenated stretched image to produce CFIs for the input of videos after applying the keyframes selection technique for the first time in the Arabic visual dataset to the best of our knowledge.

3) Investigating transfer learning different models of deep learning, Visual Geometry Group (VGG19), and VGG19 combined with Batch Normalization (BN).

The paper is structured as follows: Section 2 summarizes the related works, while Section 3 presents the existing visual speech datasets. Section 4 outlines the proposed Arabic visual speech recognition model and its stages and the collected dataset in detail, and Section 5 presents the experiments and results discussion. Finally, the conclusion and feature work are presented in Section 6.

2 Related Works

In this section, we summarize recent research conducted in lip-reading recognition techniques. Much research has been done in the field for English, French, and Chinese, as well as several other languages. In contrast, few works have attended to develop a lip-reading model for the Arabic language, and there is no public Arabic visual dataset for lip reading. Most of the reviewed lip-reading models are composed of three primary steps: pre-processing, feature extraction, and visual speech recognition. The reviewed studies are divided into two main categories: handcrafted features-based models and deep learning features-based models.

2.1 Handcrafted Features Based Models

Reda et al. [9] have created a hybrid voting model for automatic Arabic digits recognition, they have been applying three techniques for visual feature detection: Speeded-Up Robust Features (SURF), a Histogram of Oriented Gradient (HoG), and Haar feature extractor. The features are passed individually into HMM to recognize the corresponding digit. The SURF, HoG, and Haar techniques are also applied for feature extraction in [10] to build the Silent password recognition framework. The model has five layers to provide Arabic digit prediction for the password matching process, and then each feature separately is fed into the HMM. The model has achieved a high detection rate and accuracy while maintaining a low False Positive Rate (FPR).

The DCT technique is used for extracting visual features as proposed by Elrefaei et al. [11], the results show that the average Word Recognition Rate (WRR) is 70.09%, by applying the SVM for classification. Also, Catalán in [12] has utilized the DCT coefficients, HMM, Gaussian Mixture Models (GMM), to build the visual password model for user authentication, by using the GRID dataset [13] to measure the performance of the models. The results show the model performed well when recognizing users, with an equal error rate of 4.23%. This approach can be considered a method for user identification rather than word recognition because it does not achieve lip-reading recognition; the word recognition errors were too high.

The multi-view audiovisual dataset has been produced for mouth motion analysis by the authors in [14]. They concentrate on building a more flexible VSR that supports a multi-view dataset called OuluVS2. In the feature extraction stage, they use 2D DCT, and PCA for feature reduction. The results show that a 60° angle provides an average of 47% as the highest recognition rate.

Liang et al. [15] have defined the automatic lip-reading model. Firstly, they have used an XM2VTS dataset. They localize the mouth by applying the Support Vector Machine (SVM), and the extraction of visual features from lip movements is executed using hybrid PCA and Linear Discriminative Analysis (LDA). All the extracted features and acoustic sequences are combined using a coupled HMM. The result of the WRR is 55%.

Komai et al. [16] have presented a model to extract the lip movement's features from various face directions, by transforming the sideways view of the lip into a frontal view. They have applied the Active Appearance Models (AAM), followed by the HMM for the recognition. The model has achieved 77% and 80.7% in average accuracy for visual recognition rates with and without normalization of the face direction, respectively. The dataset is 216 words in the Japanese language.

Luettin et al. [17] have utilized a statistical model called an Active Shape Model (ASM) for modeling the mouth shape, which is then applied to perform the localization, parameterization, and tracking of the lip movements. They have presented two types of lip models: the first one for the outer lip contour and the second for the representation of both inner and outer lip contour. The recognition phase was done by applying HMM. The experiments were performed by utilizing the Tulips1 dataset [18], which was the first datasets used in this field. The experiments proved that the accuracy of a single contour model is 80.21 % and 80.42% for the double contour model.

Yavuz et al. [19] have implemented a lip-reading model where the features extracted by applying the PCA. They extracted the features of inner width, outer width, height between outer lips, height between inner lips, and distance between the lip, and peaks of lips. They have applied dynamic thresholding to locate the lip. A manifold representation performs the visual lip movement recognition by presenting a generic framework to recognize the words. They have tested the proposed model using 56 different videos, including Turkish vowels, and the results prove the accuracy to be 60%.

Hazen et al. [20] have extracted the features by applying PCA from visual lip measurements. For recognition, they have applied a segment constrained HMM. The collected AV-TIMIT dataset consists of 450 English sentences. The results clarify that there is a reduction in phonetic error rate with 2.5%.

Shaikh et al. [21] have presented a model to classify discrete utterances. They have calculated the optical flow (statistical properties) vectors, and the SVM then performed the classification. They have achieved an accuracy of 95.9%.

Ibrahim et al. [22] have presented a geometrical-based automatic lip-reading. The geometric feature-based approach, called a Multi Dimension Dynamic Time Warping (MDTW) uses a skin color filter, a border following method, and the calculation of convex hulls, to provide recognition of the digit. The model achieved 71% in word recognition accuracy for the CUAVE dataset [23].

Sharma et al. [24] have investigated audio-visual speech recognition for the Hindi language by utilizing the Mel Frequency Cepstral Coefficient (MFCC) for audio features. For visual features, they have used the Optical Flow (OF) and conventional lip localization approach, and then it is followed by a HMM for recognition. The dataset is 10 numeral digits from Hindi, prepared by recording videos of 24 speakers in which each speaker pronounces each digit 10 times. The 240 samples were divided for training and testing into 200, and 40 samples. The WRR is 93.76%.

The system proposed by Sagheer et al. [25] applies the hyper column neural network model combined with HMM to extract the relevant features. The system is considered the first to be

applied to VSR in the Arabic language. The experiments used a dataset of nine Arabic sentences. They compared Arabic language recognition results with those of the Japanese language, with the performance for the Arabic dataset recorded as 79.5% and 83.8% for Japanese.

2.2 Deep Learning Features Based Models

Lu et al. [26] have proposed lip-reading architecture that combines CNN with a Recurrent Neural Network (RNN) with an attention mechanism. They applied the LSTM network for retrieving hidden information in a time series. CNN was enhanced by using the VGG19 model, achieving 88.2% of accuracy.

Zhang et al. [27] have implemented an architecture model for visual-Chinese lip-reading called LipCH-Net. They applied concepts from deep neural networks model: CNN with LSTM and Gated Recurrent Unit (GRU) to recognize lip movement. The architecture combines the two models because the Chinese language needs to achieve recognition for Picture to-Pinyin (mouth motion pictures to pronunciations) and Pinyin-to-Hanzi (pronunciations to texts). They gathered daily news from Chinese Central Television (CCTV). The proposed model has accelerated training and overcome ambiguity in the Chinese language. LipCH-Net accomplishes 50.2% and 58.7% accuracy in sentence-level and Pinyin-level tasks respectively.

The main contributions of the proposed model in [28] are building concatenated stretched lip images for the input video in order to execute a deep-learning model for lip reading, by applying a 12-layer CNN with two layers of batch normalization. The validation of the performance was on MIRACLE-VC1 dataset [29]. The results show 96% in training accuracy and 52.9% in validation accuracy.

Wen et al. [30] have stated that a model with few parameters will decrease performance complexity. They have applied LSTM for extracting the sequence information between keyframes. The collected dataset represents the pronunciation of 10 English digit words (from 0 to 9). The accuracy of the model is 86.5%, which indicates that the model saves computing resources and memory space.

Faisal et al. [31] intended to combine both models to enhance speech recognition in load environments; however, they were unable to merge both networks to confirm the results of audio-visual recognition. They have applied two different deep-learning models for lip-reading spatiotemporal convolution neural network and Connectionist Temporal Classification Loss (CTC), and for audio, they have used the MFCC features in a layer of LSTM cells and output the sequence for the Urdu language.

Petridis et al. [32] have presented an end-to-end visual speech recognition model by applying LSTM networks. The model has two streams: one for features extraction from the pixel values of the lip region, and the other is called a different image by computing the difference between two sequential frames. The second streams utilized a Bidirectional LSTM (Bi-LSTM) with a pre-trained model of Restricted Boltzmann Machines (RBMs), after which the softmax produces the predicted result. The OuluVS2 [14] and CUAVE datasets [23] were used for testing and validating with a classification accuracy of 84.5% and 78.6%, respectively.

According to [33], with the variety of models and methods that have been proposed, an overall lip-reading model is produced by applying six type of models: a CNN with LSTM baseline model, a deep layered CNN with LSTM model, an ImageNet pre-trained VGG-16 features with LSTM model, and a fine- tuned VGG-16 with LSTM model. For expanding the number of training sequences of the MIRACL-V1 dataset size, they have appended horizontally flipped and

pixel-jittered versions of each image. The best accuracy of 79% was achieved with the fine- tuned VGG with LSTM.

Assael et al. [34] have applied a spatiotemporal CNN with Bi-LSTM and CTC. It is the first application to work in the end-to-end model for sentence-level. The spatiotemporal convolutions with GRUs were applied, and CTC loss function was used for training. The GRID dataset [13] was employed for evaluating, which indicated the highest accuracy to be 95.2%. A more profound architecture is used by Stafylakis et al. [35], they have proposed a model consisting of the following components: the front-end, which is the spatiotemporal convolution to the frame sequence, a Residual Network (ResNet), and the backend two-layer of Bi-LSTM. The softmax layer then provides word recognition from the LRW dataset (BBC TV broadcasts), attaining 83% accuracy.

Chung et al. [36] have trained the ConvNet to identify the visual features which passed as an input to LSTM for 10 phrases from the OuluVS2 [14] dataset (frontal view). The proposed work concentrated on combining the sound and mouth images as a synchronization, as well as a non-lip-reading problem statement. The achieved high accuracy in the LSTM based model with the SyncNet pre-trained model.

Garg et al. [37] have used images of celebrities from Google Images and IMDB for training the pre-trained VGGNet model. For extracting the features, they applied two approaches: the LSTMs for a sequence of frames, and the CFI for each video. The dataset used is MIRACL-VC1 [29]. The model achieved best accuracy in CFIs of 76%.

Mesbah et al. [38] have focused on the application of this technology to the medical field, particularly for people who suffer from laryngectomies. The model consists of the Hahn Convolutional Neural Network (HCNN). The first layer of Hahn moments eliminates the video image dimensionality and provides the extracted features to CNN. The training was executed using different datasets: AVLetters, OuluVS2, and BBC LRW.

Wang et al. [39] have sought the solution for sign language word classification, to provide a bidirectional communication system for sign language and visual speech recognition. For the feature extraction phase in their VSR, they applied the ConvNet unit (DenseNet) and Bi- LSTM as well as CTC loss to identify the recognized sentence. The dataset utilized was the public data LRW for the Chinese language, 49000 videos were used for training and 500 videos for testing. They have reached an average recognition rate of 35%.

Bi et al. [40] have developed a DenseNet network structure, the 3D convolution neural network with LSTM (E3DLSTM), which handles the time modeling for features extraction. The CTC layer is then utilized as a cascading time classification. The results show a high recognition rate compared with traditional methods for the Chinese language. The proposed model achieves the following accuracy results according to a length of letters corresponding to pinyin: 38.96% for easy, 38.49% in medium and 37.92% in hard.

Saitoh et al. [41] have built a model that depends on the concept of CFI for the required pre-processing of video frames with two types of dataset augmentation, with CNN applied for features extraction. For classification, the softmax layer with cross-entropy loss has been utilized. They have employed the OuluVS2 dataset (frontal view) to evaluate the method using different pre- trained models. The accuracy is 61.7% with the NiN model and 89.4% with the GoogLeNet model.

Tab. 1 summarizes the existing studies whose three main stages we have discussed: preprocessing, features extraction, classification, and finally the results are presented.

Table 1: Summary of lip-reading existing techniques.

	Reference# (year)	Preprocessing method		Feature extraction method	Recognition method	Language	Dataset	Results
		Face recognition	Lip region					
Handcrafted features based models	[9] (2017)	Viola jones algorithm	Haar corner detection technique	SURF, Hog, and Haar feature extractor.	HMM with 10 hidden states with voting schema	Arabic	2000 records of the ten digits	Accuracy = 96.2%
	[10] (2020)	Viola jones algorithm	Haar corner detection technique	SURF, Hog, and Haar feature extractor.	HMM with 10 hidden states with voting schema	Arabic	2000 records for the ten digits	Accuracy = 96.2%
	[11] (2019)	N/A	manually cropped for mouth regions	DCT	SVM	Arabic	1100 videos for 10 Arabic words	Word Recognition Rate (WRR) = 70.09%
	[12] (2018)	Face landmarks (face points extracted by the tracker)	Find Mouth region (rectangle of a frame, then mouth snipped)	DCT	HMM-GMM	English	GRID [13]	Accuracy = 73.1
	[14] (2015)	SURF features	facial landmark	DCT and PCA	HMM	English	Digits, phrases, and TIMIT sentence	WRR = 47%
	[15] (2002)	The face detection algorithm	SVM	PCA and LDA	Coupled Hidden Markov Model (CHMM)	English	XM2VTS dataset (700 sequences and 139 digits)	WRR = 55%
	[16] (2012)	AdaBoos, using the Haar-like features	locate the lip area for all face directions	Active Appearance Models (AAM).	HMM with 5 hidden states	Japanese	216 words	Accuracy for single regression = 78.67%, Accuracy for multiple regression = 79.5%
	[17] (1996)	N/A	Active Shape Model (Two models of the lips)	PCA	HMM	English	Tulips1 dataset [18]	Best Accuracy were for double contour model = 80.42%
	[19] (2008)	skin color determination algorithms	Color space Analysis (Extract lips ROI in HSV)	PCA	Generic framework to recognize words without resorting to Viseme classification	Turkish	56 different videos for vowels	Accuracy = 60%

(Continued)

Table 1: Continued

	Reference# (year)	Preprocessing method		Feature extraction method	Recognition method	Language	Dataset	Results
		Face recognition	Lip region					
	[20] (2004)	The face detector	Mouth tracker, AVCSR toolkit (SVM)	PCA and LDA	HMM	English	450 TIMIT-SX sentences	2.5% reduction in phonetic error rate
	[21] (2010)	The recorded of the videos were on mouth region		Statistical properties of vertical Optical flow component.	SVM	English	Each speaker recorded 14 phonemes	Accuracy = 95.5%
	[22] (2015)	Viola Jones algorithm	Skin detection method	MDTW	novel template probabilistic MDTW	English	CUAVE [23]	WRR = 71%
	[24] (2019)	N/A	Binarization, like color intensity mapping and Pseudo-Hue methods	Mel frequency cepstral coefficient , optical flow features and conventional lip localization approach	HMM	Hindi	Ten Indian digits	WRR = 93.76%
	[25] (2004)	N/A	N/A	Hyper column neural network Model	HMM with 5 hidden states	Arabic	Nin sentences (26 words in total)	Accuracy of Arabic language = 79.5%, Accuracy of Japanese language = 83.8%
Deep learning features based models	[26] (2019)	OpenCV library	Dlib toolkit to determine the seven keys' points of the mouth	CNN with attention-based LSTM, pre-trained model (VGG19)	SoftMax layer, Two fully connected Layers	English	3000 recorded videos for DIGITS (from 0 to9)	Accuracy = 84%
	[27] (2019)	API of OpenCV	By using the facial landmark to crop the lip region with size of 120*120.	CNN with LSTM/Gated Recurrent Unit (GRU), pre-trained Model (ConvNet)	CTC, CET and SoftMax layer	Chinese	20,495 sentences from the CCTV website	average Pinyin level accuracy = 50.2%, Average sentence level accuracy = 50.2%

(Continued)

Table 1: Continued

Reference# (year)	Preprocessing method		Feature extraction method	Recognition method	Language	Dataset	Results
	Face recognition	Lip region					
[28] (2018)	The face detection model	Haar Cascade Facial Landmark, CFI are generated	twelve-layer CNN with 2 layer of batch normalization	SoftMax layer	English	MIRACL-VC1 [29]	Validation accuracy = 52.9%, training accuracy = 96.5%
[30] (2019)	Multi-Task Cascade Convolution Network (MTCCN) for face detection and correction, it is also employed to extract lip region		LSTM with attention-based, Pre-trained model (MobileNets)	SoftMax layer	English	6000 records (ten digits from 0 to 9)	Accuracy = 86.5%
[31] (2018)	Viola Jones algorithm	N/A	Spatiotemporal convolution neural network, Bi-gated recurrent neural network and CTC Loss	CTC Loss, SoftMax layer	Urdu	ten words and ten phrases	Accuracy = 72%
[32] (2017)	N/A	bounding box to extract the mouth ROI	Bi-LSTM, pre-trained Model is Restrict Boltzmann Machines (RBM)	SoftMax layer	English	OuluVS2 [14] and CUAVE [23] datasets	OuluVS dataset = 84.5%, CUAVE dataset= 78.6%
[33] (2017)	Python facial recognition library, dlib with OpenCV		CNN +LSTM baseline model, Pretrained VGG-16 Features + LSTM model, Fine-Tuned VGG-16 +LSTM model.	SoftMax layer	English	MIRACL-VC1 [29]	Accuracy = 59%
[34] (2015)	DLib and the iBug face predictor	Affine Transformation	Spatiotemporal convolution neural network followed by Bi-gated recurrent neural network (LSTM)	CTC Loss and SoftMax layer	English	GRID [13]	Accuracy = 95.2%

(Continued)

Table 1: Continued

Reference# (year)	Preprocessing method		Feature extraction method	Recognition method	Language	Dataset	Results
	Face recognition	Lip region					
[35] (2017)	N/A	Focused on mouth region by applying the heatmaps extraction per landmark	Spatiotemporal convolution neural network (a 3D convolutional front-end, a ResNet and an LSTM-based back end)	SoftMax layer	English	LRW (from BBC TV broadcast)	Accuracy = 83 %
[36] (2016)	The HOG-based face detection method	N/A	MFCC features and uni-directional LSTM with 250 hidden units, Pre- trained models: VGG and SyncNet	SoftMax layer	English	OuluVS2 [14] (frontal view)	The best accuracy with SyncNet pre trained model: 94.1% for short phrases and 92.8% for digit
[37] (2016)	OpenCV library	N/A	CNN for CFI, and LSTM for videos, Pre-trained model is VGG19	SoftMax layer	English	MIRACL-VC1 [29]	Accuracy = 76%.
[38] (2016)	N/A	N/A	Hahn moments and CNN	SoftMax layer	English	AVLetters, OuluVS2 and LRW	The accuracy is 93.72% for OuluVS2, 90.86% for LRW, and 59.23% for AVLetters
[39] (2019)		N/A	ConvNet Unit (DenseNet) and Bi- LSTM	CTC	Chinese	LRW	Average recognition rate = 35%.
[40] (2019)		N/A	DenseNet network structure and E3D-LSTM is adapted with Lookahead optimizer	CTC	Chinese	LRW	The accuracy according to a length of letters: for easy is 38.96%, 38.49% in medium and 37.92% in hard.
[41] (2016)		N/A	CFI-based CNN	SoftMax with cross-entropy loss	English	OuluVS2 [14] (frontal view)	Best recognition with GoogLeNet model = 91.7%

3 Visual Speech Datasets

The visual speech datasets vary in terms of language, acquisition device, number of speakers, resolution, number of frames, size, speed of speech, and many other characteristics [11]. Several

public visual datasets are available to visual speech recognition researchers for the purpose of development in languages other than the Arabic language, such as GRID [13], OuluVS2 [14], CUAVE [23], and MIRACL-VC1 [29]. Since there are no public Arabic datasets, we have collected our dataset as will be explained in section 4.1. Tab. 2 summarizes the existing datasets that have been collected by researchers in the field.

Table 2: Summary of audio-visual speech datasets

Reference # (year)	No. of speakers	Resolution and frame rate	Language	Utterances	Public
[9] (2017)	20 speakers (13 male and 7 females	640 × 480 pixel, 30 fps	Arabic	2000 records for the ten Arabic digits (0 to 9)	No
[11] (2019)	22 speakers (8 male and 14 female)	1920 × 1080 pixel, 30 fps	Arabic	1100 utterances for 10 daily communication Arabic words	No
[13] (2006)	34 speakers	720 × 576 pixel, 25 fps	English	1000 utterances of a series of words from a 51 words vocabulary	No
[14] (2015)	53 distinct speakers (40 males and 13 females)	1920 × 1080 pixel, 30 fps	English	records 1000 of daily phrases	Yes
[16] (2012)	N/A	320 × 240 pixel, 30 fps	Japanese	216 Japanese words	No
[18] (1995)	12 speakers (9 males and 3 females)	100 × 75 pixel, 30 fps	English	first four English digits (each digit is spoken twice)	Yes
[19] (2008)	56 speakers	N/A	Turkish	videos including Turkish vowels	No
[20] (2004)	223 speakers (117 male and 106 female)	720 × 480-pixel, 30 fps	English	450 sentences, each speaker read 20 sentences	No
[21] (2010)	7 speakers (4 males and 3 females)	240 × 320 pixels, 30 fps	English	Each speaker recorded 14 phonemes in English language	No
[23] (2002)	30 speakers	720 × 480-pixel, 29.97 fps	English	Isolated or connected Numerals.	Yes
[24] (2019)	24 speakers (6 males and 18 females)	30 fps	Hindi	The ten numeral digits of Hindi language	No
[25] (2004)	9 speakers	160 × 120 pixel	Arabic	9 different Arabic sentences	No
[26] (2019)	6 speakers(3 male and 3 female)	1920 × 1080 pixel, 25 fps	English	English numbers from 0 to 9(each word pronounced up to 100 times)	No
[27] (2019)	N/A	N/A	Chinese	20,495 Chinese sentences from the CCTV website Unconstrained sentence-level	Yes
[29] (2014)	15 speakers (5 male and 10 female)	640 × 480 pixel	English	3000 recorded videos (10 words and 10 phrases)	Yes
[30] (2019)	6 distinct speakers (3 male and 3 female)	1920 × 1020 pixel, 25 fps	English	6000 records for the ten English numbers from 0 to 9	No
[31] (2018)	10 speakers	N/A	Urdu	10 words and 10 phrases of Urdu language.	No

4 Proposed Arabic Visual Speech Recognition Model

In this section we define four major stages of our proposed VSR model: data collection, data preprocessing, feature extraction, and classification as presented in Fig. 2. For any uttered word, there is a different accent. Therefore, we extract a series of images that represent the lip movement. All this temporal information will be produced into one single spatial image, as we will discuss in the next sections.

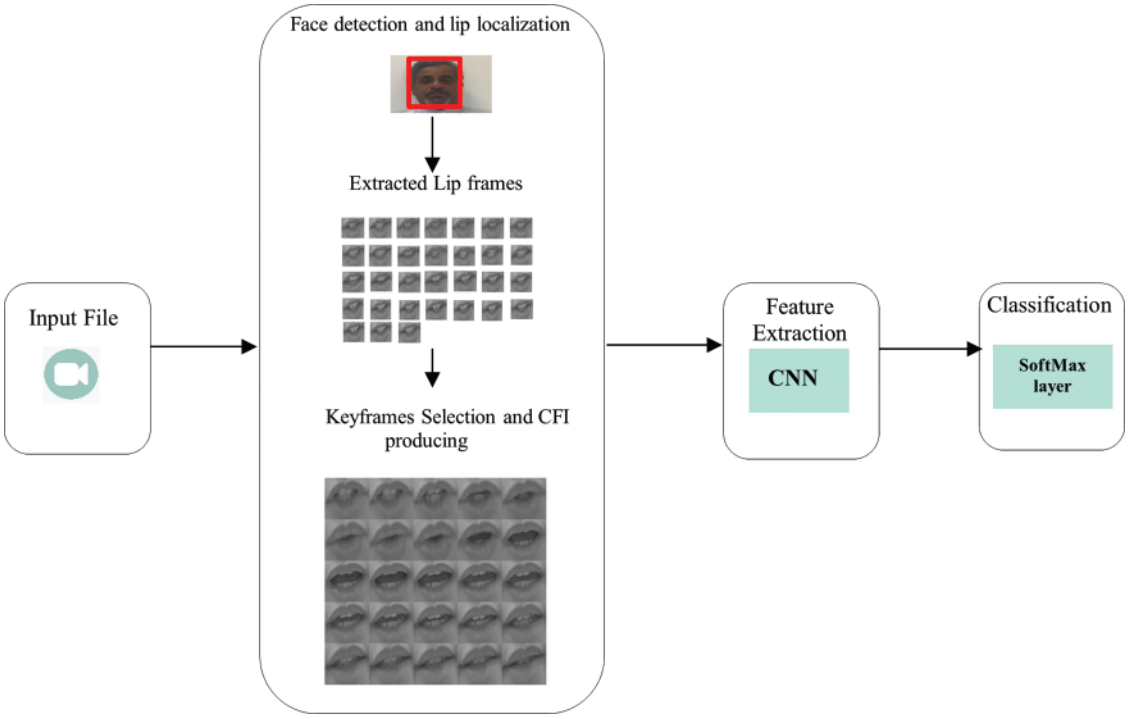

Figure 2: The proposed model architecture

4.1 Collected Dataset

The public visual datasets are available to researchers in languages other than Arabic. Since no public Arabic datasets are available, we decided to collect the required data by considering all possible difficulties such as background, speed of speech, speaker posture, and the lighting. The collected experimental dataset for the proposed model is based on 10 independent Arabic digit utterances (from 0 to 9) and 4 selected Arabic phrases. The resolution of each frame is 1920 × 1080 and approximately 25 fps, using a smartphone camera with a ring light stand. The number of speakers is 24 native speakers (14 males and 10 females). Each speaker pronounces 10 Arabic numerical words and 4 phrases, and each number and phrase repeated 10 times, resulting in 2400 records for digits and 960 records for phrases. Data acquisition was done among one month and produced with white backgrounds, and there were male speakers with and without a mustache or beard. The dataset used is shown in Tabs. 3 and 4. To accurately determine the start and end of each recorded video, we have tracked the audio to separate each spoken word. The recorded video is approximately 1 second for each digit and 2 seconds for each phrase. After preparing each video and performing frame extraction, we produced a concatenated image of 224 × 224

pixels, which contains a sequence of the uttered number or phrase as standard inputs of the CNN model.

Table 3: Dataset of selected Arabic phrases

Phrase	Arabic Pronunciation	English Translation	Number of samples
السلام عليكم	Assalam ealaykum	Peace be upon you	240
اتصل بالاسعاف	Atasil bialaiseaf	Call an ambulance	240
اتصل بالشرطة	Atasil bishurta	Call the police	240
احتاج للمساعدة	Ahtaj lilmusaeada	I need a help	240

Table 4: Dataset of Arabic digits

Digit	Arabic Pronunciation	English Translation	Number of samples
صفر	Sefr	Zero	240
واحد	Wahed	One	240
اثنان	Ethnan	Two	240
ثلاثة	Thlatha	Three	240
أربعة	Arbaa	Four	240
خمسة	Khamsa	Five	240
ستة	Setta	Six	240
سبعة	Sabaa	Seven	240
ثمانية	Thamania	Eight	240
تسعة	Tesaa	Nine	240

4.2 Data Preprocessing

The preprocessing stage consists of three main steps: face detection and lip localization, keyframe selection from lip frames, and finally the producing CFIs. As discussed in the following points.

- **Face Detection and Lip Localization:** The OpenCV library has been employed for loading the images and transforming them into a 3-dimensional matrix [21]. Then the facial landmark detection is employed from the Dlib toolkit, by using the face images as inputs for the method, in order to return the 68 landmarks of the face structure in these images. Then to crop the region of the mouth, we used the landmarks to locate the key points of the mouth region (landmarks between 48 and 68), to decrease the complexity of repetitive information and computational processing [26].

- **Keyframe Selection**: After localizing the lips and producing cropped frames of the lip region, we applied a keyframes selection technique to remove redundant information from the extracted lip frames sequence. The many factors that vary in the recorded videos, such as the length, make the process more difficult for the model to extract hidden features through all sequences. The applied keyframes extraction technique depends on calculating the average pixel intensity of the two consecutive frames [42]. As illustrated in Algorithm 1, the absolute difference between two consecutive frames is computed. Then we get the average inter-frame difference intensity by dividing the sum of the absolute difference by 64 × 64 pixels, which is the size of each lip frame. In the end, we select the frames with the largest average inter-frame difference as keyframes. We performed our experiments to select the appropriate keyframes by testing different numbers of frames: 10, 15, and 20 for Arabic digits and phrases.
- **Concatenated Frame Images**: In this step, we stretch the keyframes' sequence to produce concatenated lip image by duplicate some frames' sequence to fit a length of $L = 25$ as shown in Fig. 3, similar to the method in [37]. The speaking speed differed from one speaker to another. Consequently, the speed was normalized by applying Eq. (1) where i indicates the index of the frame in the CFI structure (from $i = 0$ to $L - 1$), *orig_images* are the selected lip frames. Fig. 4 clarifies the overview structure of CFI with keyframe = 20, as highlighted five frames out of 20 have been duplicated to normalize the speaking speed. For each row and column with the image size 320 × 320 pixels, five lip images were concatenated, and each image has a size of $w' \times h'$ [pixels], where $w' = 64$ and $h' = 64$. In the end, a total of 25 lip images were represented in one single image, which was resized to the fixed dimension of 224 × 224 pixels, as standard input to VGG19 in our architecture.

$$seq_{image}[i] = orig\_images\left[int\left(\frac{i * orig_{length}}{25}\right)\right] \tag{1}$$

Algorithm 1: The difference frames order technique

//Input: List of Lip_frames, num_top_frames\\list of all lip frames number of keyframes
//Output: Lip_keyframes.

1: Initialize the frame index $i \leftarrow 0$, *frame_shape* = 64 × 64
2: **FOR each** *frame* in *Lip_frames* **DO**
  2.1: *current_frame* ← *frame*
  2.2: **IF** (*current_frame* is not None and *prev_frame* is not None) **THEN**
    4.1: *absolute_ difference* ← (|*current_frame* − *previous_frame*|)
    4.2: *difference _sum* ← ∑ (*absolute_ difference*)
    4.3: *difference_sum_avg* ← *diff _sum*/(*frame_shape*)
    4.4: *frames_list* ← (*diff _sum_avg*, *i*)
  ENDIF
  2.3: *previous_frame* ← *current_frame*
  2.4: Increment *i* value.
3: **ENDFOR**
4: Select The first *num_top_frames* with the largest *difference_sum_avg*
5: Return the selected Lip_keyframes.
6: **END**

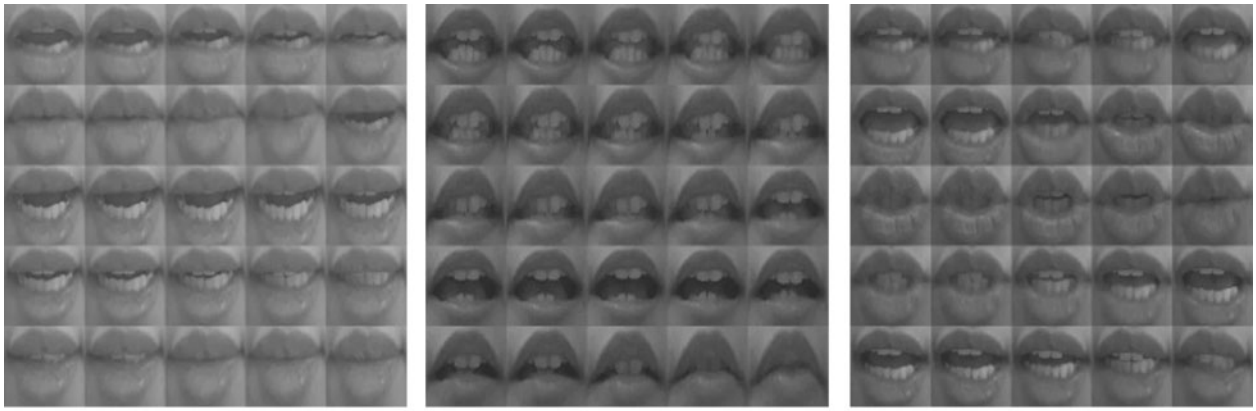

Figure 3: Concatenated Sample Dataset Images (key frames = 20)

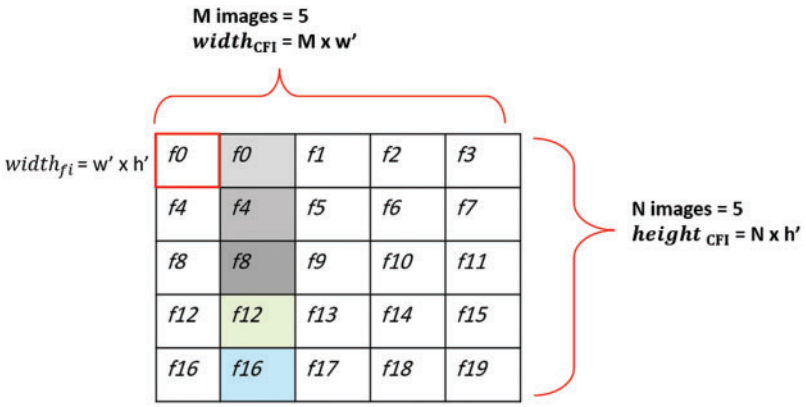

Figure 4: Overview Structure of Concatenated Frame Images (keyframes = 20)

4.3 Feature Extraction

Deep learning algorithms are adopted to extract deep features to learn long-term dependency information. We have utilized the concept of convolutional neural networks with different techniques: transfer learning and batch normalization as presented in next sections.

- **Transfer Learning**: The efficient and active way for researchers to develop and solve a problem is employing transfer learning by employing the pre-trained networks and utilizing a similar problem that the network has been trained on ImageNet [43]. The most commonly used pre-trained model is VGG, and we chose to use the technique of transfer learning by applying a VGG-19 network which is pre-trained on ImageNet to improve the part of CNN. Fig. 5 define the architecture of VGG-19. In our model, the VGG-19 does not include the last two fully connected layers, and we continued training the model based on the pre-training parameters of a grey-scale image with the input size of 224 × 244 pixel.

- **The Batch Normalization (BN)** technique is usually applied in a deep neural network to perform the network's training more efficiently, by stabilizing the training process, avoiding overfitting problem, and improve the performance of the model [44], it added after a convolutional layer in the model.
- The following points describe the proposed approaches in our architecture:

  Approach 1: We examined the VGG-19 base model as shown in Fig. 5, without any additional layers, we just replace the last fully connected layer with a new one with the number of our classes in the dataset.

  Approach 2: We removed the top classification layers of the VGG19 model with unfreezing all layers, and the 2D convolution layer added with the Relu activation function. We then added the batch normalization layer followed by two Fully Connected Layers (FCs) with Relu: FC (512) + ReLu + FC (256) + ReLu, we have examined different features size in FC layers as demonstrated in [45], and the best results were with sizes of 512 and 256 as will be shown in section 5. Finally, the SoftMax layer was added for classification. Fig. 6 illustrates the concept.

Figure 5: VGG19 network structure

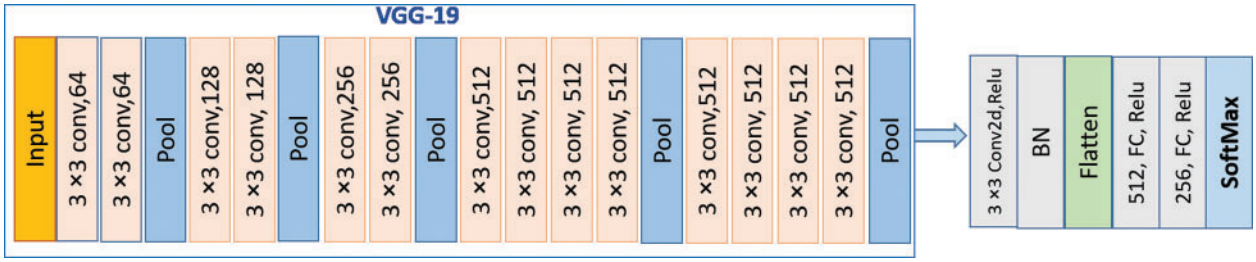

Figure 6: Approach two of fine-tuning VGG19 with batch normalization

4.4 Classification

Deep learning has made a vital rule in terms of classification accuracy. After the phase of learning the sequence of features which interprets the input video with the mouth area of a speaker saying one specific word, this sequence is employed to provide the target prediction of classification among several possible words in order to produce distribution on the class labels. Hence, the 3D representation (width × height × depth) is transformed into a vector, and finally, the vector is treated as the input to the fully connected layers and softmax convert the prediction results into probability. In our experiment, we applied the softmax with the categorical-cross-entropy to perform the classification among classes to predict the results, and the gradient descent

with Adam optimizer was used to update the weights for optimization. The SoftMax function is defined in Eq. (2), where n is the number of classes, and the exponential function was applied to the input vector $x_j$and then divided by the sum of all output vectors to ensure that the output will be equal to 1. f(x) represents the value of the function which will be between [0,1].

$$f(x_j) = \frac{e^{x_j}}{\sum_{i=1}^{n} e^{x_i}} \tag{2}$$

5 Experiments and Results Discussion

In this section, we describe the hardware and software tools, and we discuss the performance results of our model. We also compared our model with the existing approaches.

5.1 Hardware and Software Tools

The building and development of the model was conducted using an ASUS laptop with a Core i7-10750H 2.60 GHz CPU and 16 GB RAM. The Keras 2.4.3 open library with a TensorFlow 2.3.1 backend in Python 3.8.6 was used to build the deep learning model. OpenCV 4.4.0 library was employed to prepare the collected dataset, and the scikit-learn 0.23 was used to measure the performance of the model.

5.2 Results Discussion

The dataset is composed of 2400 videos for Arabic digits and 960 videos for Arabic phrases, totaling 3360 videos. The dataset was split into 3 sets: 75% for the training set, 15% for validation, and 10% for testing. We trained our model on the training set and evaluating the accuracy of the unseen data. We applied checkpoints in the training process to monitor the loss metrics by applying the callback function during the training. By monitoring the minimum loss value, the model weights were saved automatically based on the quantity being monitored. The speed of speech is different from one speaker to another, so there were a different number of frames for each recorded video. We computed the average number of frames for each class as depicted in Fig. 7. The average number of frames for digits is 33 and for phrases, it is 56. According to the variations, the production of CFIs was executed with different numbers of keyframes: 10, 15, and 20. In the experiments, we examined all discussed approaches for: digit recognition and phrases recognition, and we ultimately mixed the digits and phrases datasets.

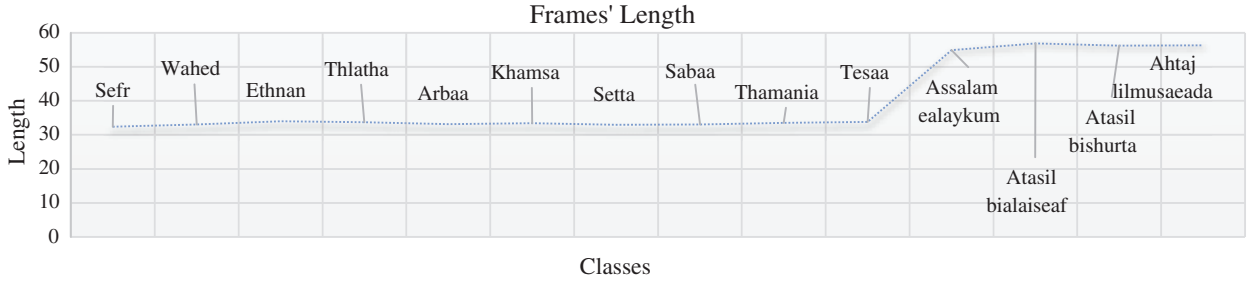

Figure 7: The average number of frame's length for each class in the dataset

Our proposed model for Arabic digit and phrase recognition has achieved the best model accuracy in the second approach with the keyframes number of 20, where the batch normalization

was utilized in VGG19 architecture. The loss has decreased and hence we can say that the digit and phrases recognition model is improving with the test accuracy of 94% for digit recognition, 97% for phrase recognition, and 93% in the experiments of digits and phrase recognition. The results are represented in Tab. 5.

Table 5: Recognition accuracy for the proposed approaches

Approach	Approach description	Number of keyframes	Arabic dataset	Training accuracy (%)	Validation accuracy (%)	Test accuracy (%)
Approach 1	VGG-19 base model (not include the top)	10	Digits	97	76	82
			Phrases	95	77	89
			Digits and Phrases	92	76	74
		15	Digits	97	77	84
			Phrases	99	84	85
			Digits and Phrases	97	76	80
		20	Digits	86	70	83
			Phrases	84.5	76	83
			Digits and Phrases	83	65	75
Approach 2	Pretrained VGG-19 (not include the top) + batch normalization layer	10	Digits	98	75	82
			Phrases	98	77	84
			Digits and Phrases	98	76	74
		15	Digits	98	76	84
			Phrases	97	87	85
			Digits and Phrases	97	78	83
		20	Digits	98.4	89	94
			Phrases	99.5	91	97
			Digits and Phrases	98	81	93

Fig. 8 gives the confusion matrices for the best model of the digit and phrase recognition experiment evaluated on test data. The matrices depicted for us the errors made, showing which digits and phrases are most often confused for one another. For example, the number 3, 'Thlatha,' is confused with number 2 'Ethnan.' This is because the speaker's mouth movements at the beginning of the utterances are the same, in order to pronounce the letter 'th.' Additionally, number 6, 'Setta,' and number 7, 'Sabaa,' are confused with number 9, 'Tessa,' as they share the same viseme sequences. The phrase 'Call the Ambulance,' or 'Atasil Bialaiseaf,' is confused with 3 digits: number 5, 'Khamsa,' number 6, 'Setta,' and number 7 'Sabaa.'

As shown in Fig. 9, in the precision and recall for digits from 0 to 9 and the 4 selected phrases, the highest precision was 100% for the number 4, which is 'Arbaa,' and 'Call the Ambulance,' which is 'Atasil Bialaiseaf.' For recall, the highest record was 100% for number 1, which is 'Wahed,' number 3 'Thlatha,' number 7, 'Sabaa,' and the phrase peace be upon you, which is 'Assalam Ealaykum.' The lowest precision appeared in number 6, 'Setta,' and number 7,

'Sabaa', and lowest recall was found in number 4 'Arbaa.' The precision and recall are calculated using Eqs. (3) and (4), where TP is True Positive predicted value, FP is False Positive where the prediction of positive value was incorrect, and FN is False Negative.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

As presented in Fig. 9, the highest F1 measure as defined in Eq. (5), recorded 92% for the phrase call the police 'Atasil Bishurta,' while the lowest F1 measure recorded 88% for the number 6 'Setta.'

$$\text{F1 Score} = 2 * \frac{(Precision * Recall)}{(Precision + Recall)} \tag{5}$$

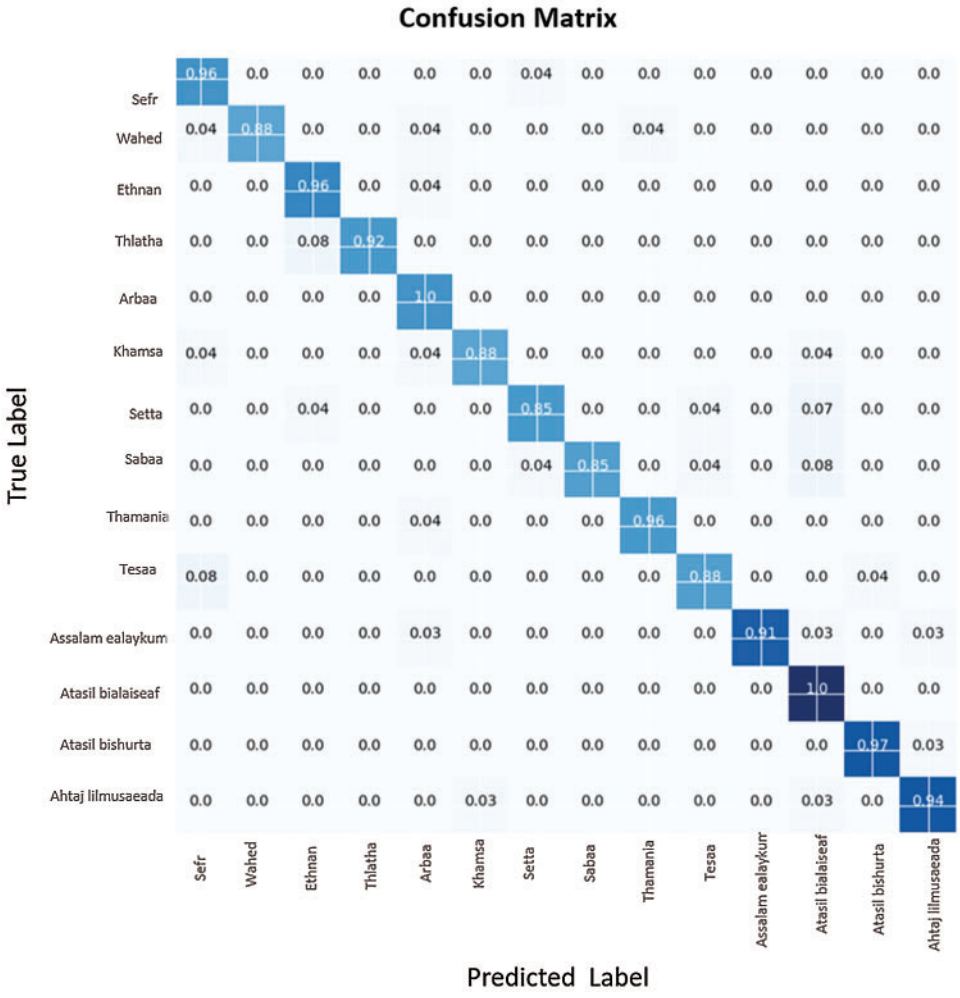

Figure 8: Confusion matrix of mixed digits and phrases model

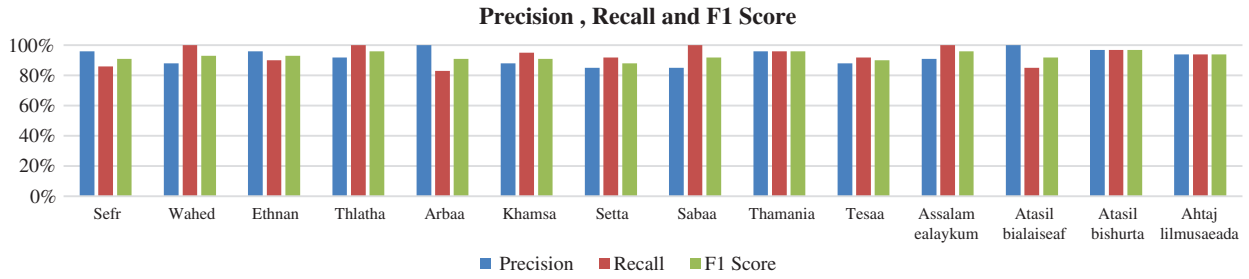

Figure 9: Precision recall and f1 score for the speech recognition model of Arabic Digits and phrases

Table 6: Comparison with the existing Arabic visual speech recognition models

Reference # (year)	Model description	Dataset	No. of speakers	Results
[9] (2017)	HMM (10 hidden states) for classification	2000 records of the ten Arabic digits	20 speakers (13 males and 7 females)	Accuracy = 96.2%
[11] (2019)	DCT with SVM for classification.	1100 videos for 10 Arabic words	22 speakers (8 males and 14 females)	WRR = 70.09%.
[25] (2004)	Hyper column neural network Model combined with HMM (5 hidden states)	Nine Arabic sentences	9 speakers	Accuracy = 79.5%
Our proposed model	Deep learning-based model, CFI- based CNN (Pretrained sVGG-19) with batch normalization)	3360 records of Arabic digits and phrases	24 speakers (14 males and 10 females)	Test Accuracy 93%
		2400 records of ten Arabic digits		94%
		960 record of Arabic phrases		97%

The experimental results show that compared with the existing models of Arabic visual speech recognition, our proposed architecture can effectively predict digits and phrases on our collected dataset: the accuracies of the proposed model are 94% for the test accuracy of digit recognition, 97% for phrase recognition, and 93% in the experiments of digit and phrase recognition. In Tab. 6, we compare our proposed model with the existing Arabic recognition model, as shown in the case of digit recognition. The model in [9] has provided an accuracy of 96.2%, and we have attained 94%. That is due to the fact that the nature of the dataset is different in terms of number of speakers according to gender, as well as the recorded video time, as the majority of their speakers were males, and the average time of the recorded videos were 30 seconds. In contrast, our collected dataset has 14 males and 11 females, and the recorded video time is one second for each digit. Furthermore, we have compared our model with state-of-the-art methods where the input of the model is concatenated frame images to the corresponding digit or phrase, as proposed in our

work. As presented in Tab. 7, our model can effectively predict the digits and phrases with the highest recognition accuracy compared to models based on CFIs input, all of which are tested with the English language.

Table 7: Comparison with models of input of concatenated frame images (CFIs)

Reference# (year)	Model description	Language	Dataset	Training accuracy (%)	Validation accuracy (%)	Test accuracy (%)
Our proposed model	CFI-based CNN (Pretrained sVGG-19) with batch normalization)	Arabic	2400 records of the ten digits	98.4	89	94
			960 record of phrases	99.5	91	97
			3360 records of digits phrases	98	81	93
[28] (2018)	CNN With Batch Normalization	English	MIRACL-VC1 dataset [29], 3000 recorded videos (10 words and 10phrases)	96	52.9	38.5
[37] (2016)	CNN with pretrained model of VGG19	English	MIRACL-VC1 dataset, 3000 recorded videos (10 words and 10 phrases)	66.15	76	44.5
[41] (2016)	CNN	English	OuluVS2 [14] dataset, 1000 records for (digits, three phrases, and three sentences)	88.5	–	–

6 Conclusion

The proposed Arabic visual speech recognition model is capable of classifying digits and phrases in the Arabic language by examining our collected visual datasets. The keyframe extraction technique and CFIs were utilized to perform the concatenated stretch image, which represents the sequence of the input video. Different approaches were used in the experiments, and the performance has been validated on our collected dataset. The results show that the CNN network with VGG19 networks to extract the bottleneck features by employing the batch normalization provides a high accuracy by stabilizing the training process, as compared to state-of-the-art methods where the input of the models is concatenated frame images. The performance of our model has yielded the best test accuracy of 94% for digit recognition, 97% for phrase recognition, and 93% in the experiments of digit and phrase recognition. Furthermore, we intend to focus on examining the model by working with different locations of the lip landmark localization by increasing the dataset to provide more accurate results in the field of visual speech recognition.

Acknowledgement: We would like to extend our thanks and appreciation to all volunteers for their participation in the dataset collecting process.

Funding Statement : The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] L. R. Aran, F. Wong and L. P. Yi, "A review on methods and classifiers in lip reading," in *IEEE 2nd Int. Conf. on Automatic Control and Intelligent Systems*, Kota Kinabalu, Malaysia, pp. 196–201, 2017.

[2] V. Sahu and M. Sharma, "Result based analysis of various lip tracking systems," in *IEEE Int. Conf. on Green High Performance Computing*, Nagercoil, India, pp. 1–7, 2013.

[3] S. Agrawal and V. R. Omprakash, "Lip reading techniques: A survey," in *2nd Int. Conf. on Applied and Theoretical Computing and Communication Technology*, Bangalore, pp. 753–757, 2016.

[4] S. Mathulaprangsan, C. Y. Wang, A. Z. Kusum, T. C. Tai and J. C. Wang, "A survey of visual lip reading and lip-password verification," in *IEEE Int. Conf. on Orange Technologies*, Hong Kong, China, pp. 22–25, 2015.

[5] N. Akhter and A. Chakrabarty, "A survey-based study on lip segmentation techniques for lip reading applications," in *Int. Conf. on Advanced Information and Communications Technology*, 2016.

[6] J. Abhishek, V. P. Namboodiri and C. V. Jawahar, "Word spotting in silent lip videos," in *IEEE Winter Conf. on Applications of Computer Vision*, Lake Tahoe, NV, USA, 2018.

[7] H. Kulkarni and D. Kirange, "Artificial intelligence: A survey on lip-reading techniques," in *IEEE 10th Int. Conf. on Computing, Communication and Networking Technologies*, Kanpur, India, pp. 1–5, 2019.

[8] D. W. Jang, H. I. Kim, C. Je, R. H. Park and H. M. Park, "Lip reading using committee networks with two different types of concatenated frame images," *IEEE Access*, vol. 7, pp. 90125–90131, 2019.

[9] A. H. Reda, A. A. Nasr, M. M. Ezz and H. M. Harb, "An Arabic figures recognition model based on automatic learning of lip movement," *Al-Azhar University Engineering Sector*, vol. 12, pp. 155–165, 2017.

[10] M. Ezz, A. M. Mostafa and A. A. Nasr, "A silent password recognition framework based on lip analysis," *IEEE Access*, vol. 8, pp. 55354–55537, 2020.

[11] L. A. Elrefaei, T. Q. Alhassan and S. S. Omar, "An Arabic visual dataset for visual speech recognition," *Proc. Computer Science*, vol. 163, pp. 400–409, 2019.

[12] P. C. Rabaneda, "Lip reading visual passwords for user authentication," B.A. Thesis, Federico Santa María Technical University, Spain, 2018.

[13] M. Cooke, J. Barker, S. Cunningham and X. SShao, "An audio-visual corpus for speech perception and automatic speech recognition," *Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[14] I. Anina, Z. Zhou, G. Zhao and M. Pietikäinen, "Ouluvs2: A multi-view audiovisual database for non-rigid mouth motion analysis," in *11th IEEE Int. Conf. and Workshops on Automatic Face and Gesture Recognition*, Ljubljana, Slovenia, pp. 1–5, 2015.

[15] L. Liang, X. Liu, Y. Zhao, X. Pi and A. V. Nefian, "Speaker independent audio-visual continuous speech recognition," in *IEEE Int. Conf. on Multimedia and Expo*, Lausanne, Switzerland, pp. 25–28, 2002.

[16] Y. Komai, N. Yang, T. Takiguchi and Y. Ariki, "Robust AAM-based audio-visual speech recognition against face direction changes," in *Proc. of the 20th ACM Int. Conf. on Multimedia*, Nara, Japan, pp. 1161–1164, 2012. https://doi.org/10.1145/2393347.2396408.

[17] J. Luettin, N. A. Thacker and S. W. Beet, "Visual speech recognition using active shape models and hidden Markov models," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing Conf. Proc.*, Atlanta, GA, USA, pp. 817–820, 1996.

[18] J. R. Movellan, "Visual speech recognition with stochastic networks," in *Advances in Neural Information Processing Systems*. vol. 7, Cambridge: MIT Press, pp. 851–858, 1995.

[19] Z. Yavuz and V. Nabiyev, "Automatic lipreading with principal component analysis," in *Second Int. Conf. on Problems of Cybernetics and Informatics*, Baku, Azerbaijan, pp. 143–146, 2008.
[20] T. J. Hazen, K. Saenko, C. H. La and J. R. Glass, "A segment-based audio-visual speech recognizer: Data collection, development, and initial experiments," in *Proc. of the 6th Int. Conf. on Multimodal Interfaces*, New York, NY, USA, pp. 235–242, 2004. https://doi.org/10.1145/1027933.1027972.
[21] A. A. Shaikh, D. K. Kumar, W. C. Yau, M. C. Azemin and J. Gubbi, "Lip reading using optical flow and support vector machines," in *3rd Int. Congress on Image and Signal Processing*, Yantai, China, pp. 327–330, 2010.
[22] M. Z. Ibrahim and D. J. Mulvaney, "Geometrical-based lip-reading using template probabilistic multi-dimension dynamic time warping," *Journal of Visual Communication and Image Representation*, vol. 30, pp. 219–233, 2015.
[23] E. K. Patterson, S. Gurbuz, Z. Tufekci and J. N. Gowdy, "CUAVE: A new audio-visual database for multimodal human-computer interface research," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Orlando, FL, USA, pp. 2017–2020, 2002.
[24] U. Sharma, S. Maheshkar, A. N. Mishra and R. Kaushik, "Visual speech recognition using optical flow and hidden Markov model," *Wireless Personal Communications*, vol. 4, pp. 2129–2147, 2019.
[25] A. E. Sagheer, N. Tsuruta and R. I. Taniguchi, "Arabic lip-reading system: A combination of Hyper-column neural," in *Proc. of Int. Conf. on Artificial Intelligence and Soft Computing*, Marbella, Spain, pp. 311–316, 2004.
[26] Y. Lu and H. Li, "Automatic lip-reading system based on deep convolutional neural network and attention-based long short-term memory," *Applied Sciences*, vol. 8, no. 9, pp. 1599, 2019.
[27] X. Zhang, H. Gong, X. Dai, F. Yang, N. Liu *et al.,* "Understanding pictograph with facial features: end-to-end sentence-level lip reading of Chinese," *Proc. of the AAAI Conf. on Artificial Intelligence*, vol. 33, pp. 9211–9218, 2019.
[28] S. H. Nadeem, H. Gupta, D. Mittal, K. Kumar, A. Nanda *et al.,* "A lip reading model using CNN with batch normalization," in *Eleventh Int. Conf. on Contemporary Computing*, Noida, India, pp. 1–6, 2018.
[29] A. Rekik, A. Ben-Hamadou and W. Mahdi, "A new visual speech recognition approach for RGB-D cameras," in *Int. Conf. on Image Analysis and Recognition*, Vilamoura, Portugal, Springer, Cham, pp. 21–28, 2014. https://doi.org/10.1007/978-3-319-11755-3_3.
[30] J. Wen and Y. Lu, "Automatic lip reading system based on a fusion lightweight neural network with Raspberry Pi," *Applied Sciences*, vol. 24, no. 9, pp. 5432, 2019.
[31] M. Faisal and S. Manzoor, "Deep learning for lip reading using audiovisual information for Urdu language," arXiv preprint arXiv: 1802.0552, 2018.
[32] S. Petridis, Z. Li and M. Pantic, "End-to-end visual speech recognition with LSTMs," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, New Orleans, LA, USA, pp. 2592–2596, 2017. https://doi.org/10.1109/ICASSP.2017.7952625.
[33] A. Gutierrez and Z. Robert, "Lip reading word classification," Technical Report, Stanford University, CS231n project report, 2017.
[34] Y. M. Assael, B. Shillingford, S. Whiteson and D. N. Freitas, "LipNet: End-to-end sentence-level lipreading," arXiv preprint arXiv: 1611.01599, 2016.
[35] T. Stafylakis and G. Tzimiropoulos, "Combining residual networks with LSTMs for lipreading," arXiv preprint arXiv: 1703.04105, 2017.
[36] J. S. Chung and A. Zisserman, "Out of time: Automated lip sync in the wild," in *Asian Conf. on Computer Vision*, Taipei, Taiwan, Springer, Cham, pp. 251–263, 2016. https://doi.org/10.1007/978-3-319-54427-4_19.
[37] A. Garg, J. Noyola and S. Bagadia, "Lip reading using CNN and LSTM," Technical report, Stanford University, CS231n project report, 2016.
[38] A. Mesbah, A. Berrahou, H. Hammouchi, H. Berbia, H. Qjidaa *et al.,* "Lip reading with Hahn convolutional neural networks," *Image and Vision Computing*, vol. 88, pp. 76–83, 2019.

[39] F. Wang, S. Sun and Y. Liu, "A bidirectional interactive system of sign language and visual speech based on portable devices," in *IEEE Int. Conf. on Robotics and Biomimetics*, Dali, China, pp. 1071–1076, 2019.

[40] C. Bi, D. Zhang, L. Yang and P. Chen, "A lip reading model with DenseNet and E3D-LSTM," in *IEEE 6th Int. Conf. on Systems and Informatics*, Shanghai, China, pp. 511–515, 2019.

[41] T. Saitoh, Z. Zhou, G. Zhao and M. Pietikäinen, "Concatenated frame image-based CNN for visual speech recognition," in *Asian Conf. on Computer Vision*, Taipei, Taiwan, Springer, Cham, pp. 277–289, 2016. https://doi.org/10.1007/978-3-319-54427-4_21.

[42] S. Vassiliadis, E. A. Hakkennes, J. S. Wong and G. G. Pechanek, "The sum-absolute-difference motion estimation accelerator," in *Proc. 24th Euromicro Conf.*, Vasteras, Sweden, IEEE, pp. 559–566, 1998. https://doi.org/10.1109/EURMIC.1998.708071.

[43] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li *et al.,* "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conf. on Computer Vision and Pattern Recognition*, Miami, FL, USA, 2009.

[44] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Int. Conf. on Machine Learning*, Lille, France, pp. 448–456, 2015.

[45] A. Al-Shannaq and L. Elrefaei, "Age estimation using specific domain transfer learning," *Jordanian Journal of Computers and Information Technology*, vol. 6, no. 2, pp. 122–139, 2020.