Mustansiriyah Journal of Pure and Applied Sciences, Print ISSN: 2957-9910. Online ISSN: 2957-9929, Vol. 3, No. 4, 2025, Pages 53-65
DOI: 10.47831/mjpas.v3i4.306

MJPAS

*MUSTANSIRIYAH JOURNAL OF PURE AND APPLIED SCIENCES*

Journal homepage:
https://mjpas.uomustansiriyah.edu.iq/index.php/mjpas

*RESEARCH ARTICLE – COMPUTER SCIENCE*

# Lip Reading to Distinguish Phrases Using Squeezenet

**Alaa Abdulraheem Yasir**
The Directorate General of Education in Dhi Qar Governorate
alaa2020_abd@utq.edu.iq

**Article Info.**

*Article history:*

Received
5 September 2024

Accepted
24 September 2024

Publishing
30 September 2025

**Abstract**

The need to build visual speech recognition is the main motive that made us research this topic. There is difficulty in understanding speech when visible, there is an urgent need to develop a system that can read lips and understand visual speech. This would help people with hearing impairments to interpret lip movements and understand the spoken phrases and sentences. It would also assist people in noisy environments such as stadiums, airports, factories, and other places where it is difficult to access audio signals. Therefore, researchers are continuously striving to find the best solutions to address this problem. This system is of great importance to empower people with hearing disabilities and others who face difficulties in understanding spoken language in noisy environments.

In an attempt to solve this problem, this research designed and implemented a real-time system to visually interpret and understand spoken phrases and sentences without the need for audio input. The proposed system consists of two main stages: the first stage is detecting the face region and the mouth region, followed by the detection and localization of the area of interest (ROI), which is the lip region. The process is carried out by capturing video of the speaker and dividing it into consecutive frames. After detecting the face region and the mouth region, the area of interest, which is the lip region, is detected. Multitask convolutional neural networks algorithm (MTCNN) was used to perform the detection and localization of these regions. The second stage involves inputting the frames corresponding to the lip region into The squeeze Net convolutional network model for recognizing the spoken phrases and sentences. The proposed method achieved an accuracy of 90%.



***Keywords:*** MTCNN; visual speech recognition; and squeeze Net.

## 1- Introduction

Speech is a highly important natural means of communication among humans[1]. In the past, the use of speech recognition techniques was limited only to the processing of audio inputs, These systems are commonly known as Automatic Speech Recognition (ASR) systems or speech-to-text systems[2]. One of the most prominent failures of speech-to-text systems is a result of several reasons, including audio source degradation, their inability to distinguish speech in the presence of background noise, or

distortions that occur in the transmission channels, among other factors[1]. As a result, there is a need for alternative sources of speech information that are not affected by audio noise. Visual speech recognition (VSR) is the most promising source in this regard, as it is unaffected by any form of audio noise[3]. Another name for visual speech recognition is automatic lip reading, lip reading has been known since approximately the sixteenth century, and it is considered one of the common means to assist people with hearing impairments, as it also makes an important contribution to understanding fluent speech[4]. There are multiple concepts associated with automatic lip reading, inclusive: a method To predict words and understand phrases solely Taken from video sans any sound cues[5]. It is also called as the process of decoding speech from the movements of the speaker's mouth[6]. Speech recognition is not considered an easy task; this is because lip reading systems face A variety of challenges caused by discrepancies in inputs, including : accents, poor lighting, Similarly, diversity of skin tones, the variation in speaking rates, facial features, and vocal intensity also pose significant challenges[7]. To mitigate these challenges, one can resort to utilizing more of the collected visual input data[5]. As in this research conducted here, 32400 images were collected for this purpose.

This paper is Ordered as follows: Section 2 literature reviews. Section 3 presents the proposed planning system. Section 4 describes the Dataset Description. Section 5 discussion of behavior tests and results. Finally, Section 6 concludes this paper.

**Nomenclature & Symbols**

| | |
|---|---|
| MTCNN | Multitask convolutional neural networks algorithm |
| ASR | Automatic Speech Recognition |
| VSR | Visual Speech Recognition |
| ROI | Region Of Interest |

## 2- Literature reviews

A.Ngzkshay et al. (2020) in [2]: The method of extracting and processing ROI from isolated letters and numbers was discussed and its effect on the accuracy of lip-reading recognition was discussed. The ROI is extracted using the well-known Viola-Jones algorithm and the features are extracted using two distinct methods LBP and DCT. Both methods do not require prior training and are therefore a good and effective way to extract features easily. The presented model achieved an accuracy of up to 83.2%.

Lu, Yuanyao et al. (2019) in [8]: The authors proposed a new approach for the ALR system that combines CNN utilized for image feature extraction with RNN built upon the attention mechanism for ALR. CNN using the model VGG19 to extracted visual features from the mouth Region Of Interest (ROI). The proposed model achieves an accuracy of 88.2%.

Befkadu Belete (2019) in [9]: An audio-visual strategy for interpreting lip movements was proposed. The method Leveraged the Viola-Jones algorithm for mouth region detection, and then employed Discrete Wavelet Transformation (DWT) to extract the features. The AAVC database was utilized in this experiment, which achieved a recognition accuracy of 72% and a discrimination rate of 67.08%.

Faisal et al. (2018) in [10]: The main goal was to improve the speech recognition in the noisy environment by combining two different models of Deep Learning : first one for video sequences using CNN, recurrent neural network (RNN), Connectionist Temporal Classification Loss, and the second for audio that inputs the Mel Frequency Cepstal Coefficients (MFCC) features to a layer of LSTM cells and output the sequence, and achieved accuracy rate 72% .

J.Chung (2017) in [11]: In the study An approach known as "spell, listen, attend, and watch" was presented. The aim of this method was to convert mouth motion videos into characters. LSTM (Long Short-Term Memory) together with CNN (Convolutional Neural Network) were used to identify the spoken words. The proposed approach achieved an accuracy of 76.2%.

A.Patil et al. (2015) in [1]: The researchers were able to perform visual speech recognition using Artificial Neural Network (ANN) and Support Vector Machine (SVM) the features used as inputs to these two methods were extracted using Discrete Wavelet Transform (DWT).These experiments were conducted using the CAUVE database, and the two methods achieved a recognition rate of 83.2%.

Zhao et al. (2009) in [12]: They implemented local spatiotemporal descriptors as their method for lip reading automation. The method involved extracting local spatiotemporal Local Binary Patterns (LBP) from the mouth region to represent the isolated words. This method was able to achieve a recognition accuracy of 58.85%.

## 3- The proposed system

The input for the suggested system is a set of video clips that were captured through the installation of cameras in front of a diverse group of participants. These participants vary in their ethnicities and features, and some of them wear prescription glasses while others have beards and mustaches. The cameras installed positioned before these participants recorded the video clips under different lighting conditions as they pronounced a number of sentences and phrases. The proposed system comprises of three main stages: preprocessing stage, Stage of using squeeze Net , and classification Stage.

**Preprocessing stage:**

The preprocessing stage is one of the crucial stages of the proposed system. It is a technique that facilitates the task of analyzing and reducing basic data by removing unnecessary data. In this stage, a number of mathematical and statistical operations are performed on the image frames. After each frame is extracted from the video, it is sent to the face detection algorithm, where the face region is considered the target in this step [13] , After that the frames corresponding to the face region are sent to the same algorithm again to extract the area of interest, which is the mouth region. These operations extract the important features from these frames and determine the region of interest (ROI) using geometric operations such as cropping, zooming, rotating, and translating. Consequently, this stage works to enhance and prepare the data for the subsequent stages of the system.

**Multi-tasking convolutional neural networks (MTCNN):**

It is one of the algorithms adopted in convolutional neural networks in detecting faces, because it has a high detection accuracy [14]. The work of (MTCNN) goes through three stages:-

The first step: The algorithm (MTCNN) creates multiple frames scanning through the image, starting from the upper corner The left and then advance towards the right lower corner and this process is called information retrieval P-Net (Proposal Net).

The second step: All information received from the network (P-Net) will be used as input for the next layer, called the transformation network R-Net (Refinement Network), which rejects most frames that do not contain faces[15].

The third step: Which is more complex and known as O-Net (Output Network),the output network [14].

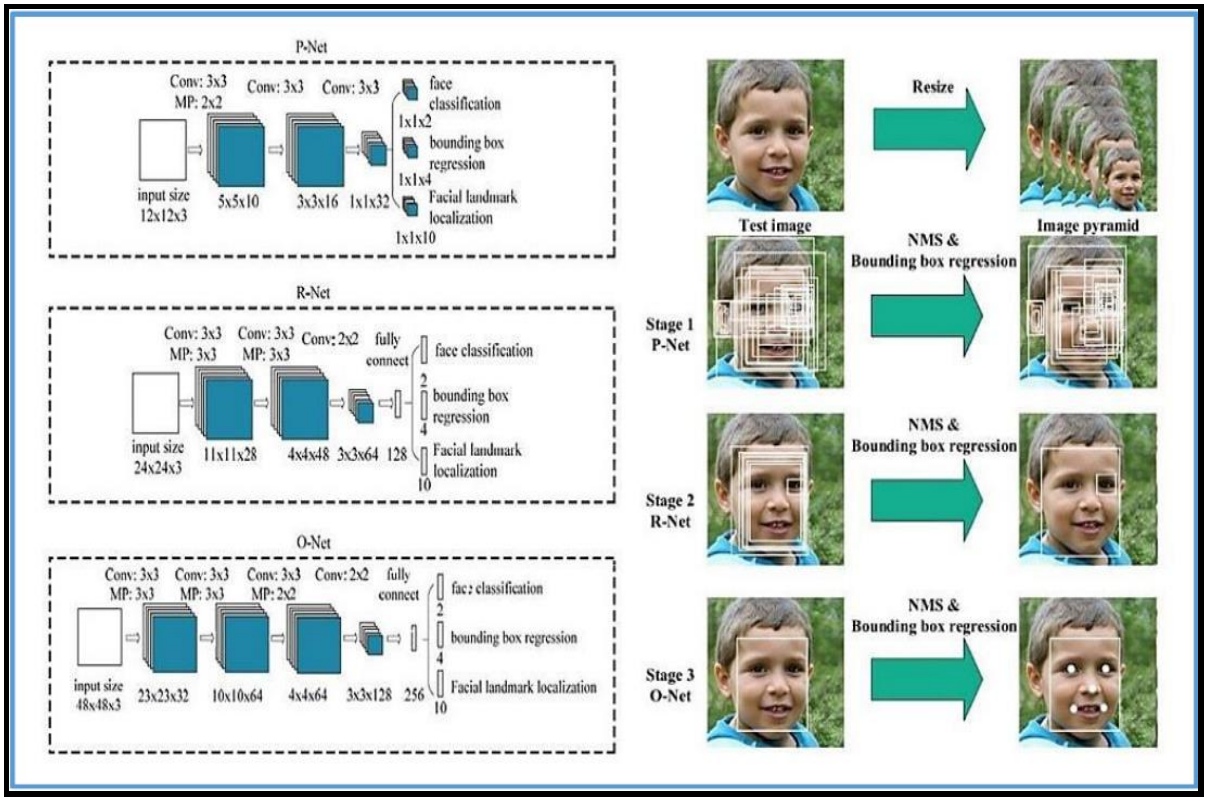


**Figure 1 :The architecture of MTCNN that used for face detection and landmark extraction**[15]

**Figure 2: Face detection using the MTCNN face detector** [15].

## Face and mouth region detection

### Face detection:

Face detection is one of the common computer vision problems that aims to locate faces within a digital image [16]. Face detection is considered an open field due to various factors such as differences in person positioning, facial expressions, scale, lighting, image distortion, face occlusion, and other elements[17]. Face recognition has been a significant challenge due to the wide variations in facial features found in photographs. Differences in face size within the image, its location, orientation (such as being rotated or not), and pose (like frontal, side, or profile) make it difficult to construct an effective face recognition system[18]. A set of algorithms have been proposed to address these challenges, with multi-task convolutional neural networks (MTCNN) being one of the most prominent approaches[6].


**Figure 3: Sample images showing different variation factors** [19].

Segmenting the video into ordered frames, followed by defining

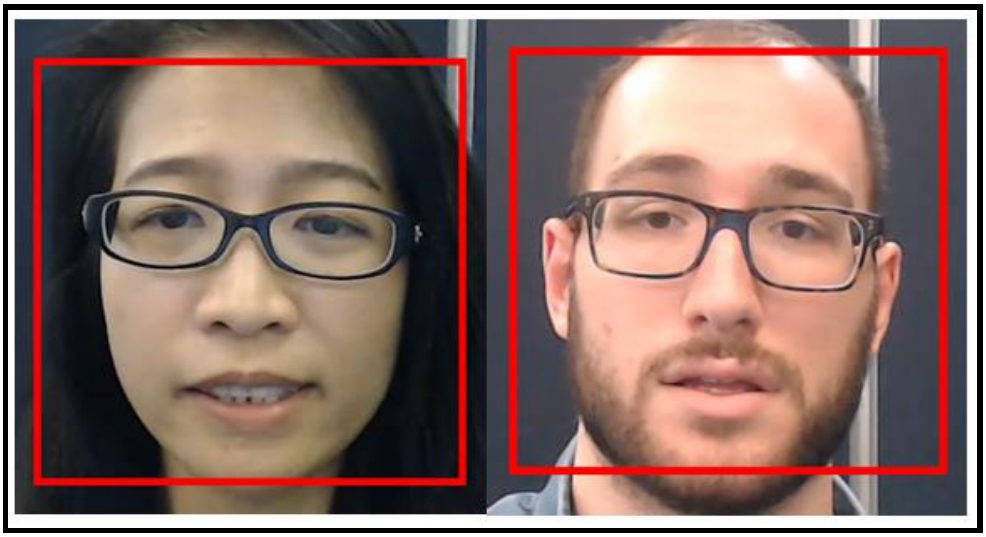

**Figure 4: Sample for face region detection MTCNN algorithm**

## Detection of the mouth

Following the capture of a video clip of Belonging to the individual pronouncing the specified words, It is saved in a temporary folder along with the facial area and then detection of the mouth is achieved by the MTCNN algorithm. As demonstrated in the figure (5) detection of the region of the face and mouth.

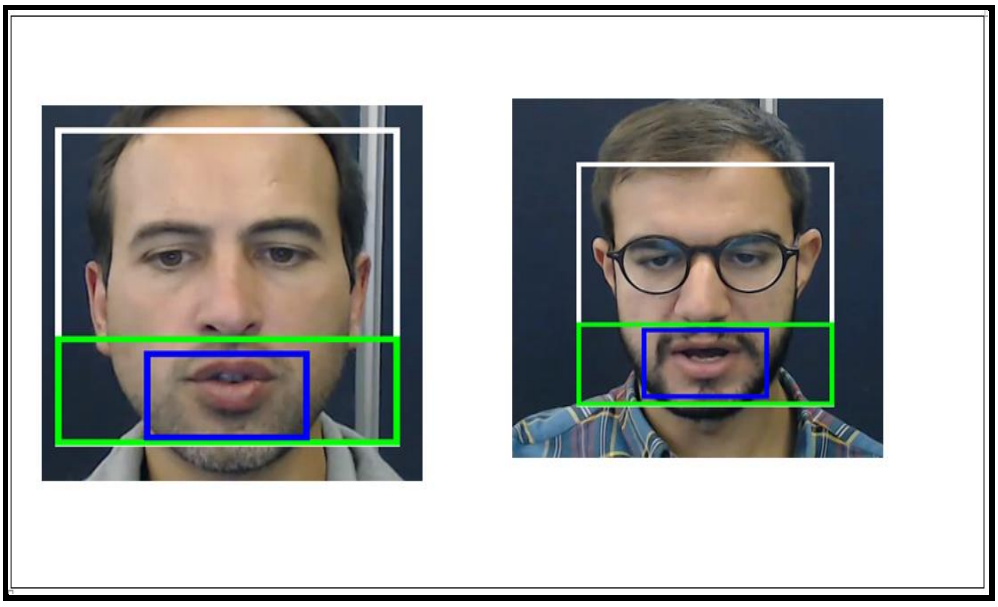

**Figure 5: Sample demonstrating the detection of the face and mouth area by MTCNN algorithm**.

## Cut the Image of the mouth extracted from the frame

The operation of image cropping involves selecting the areas of interest (ROI) in the image, which are named a sub-image. This process helps us in involvement in the analysis stage. At this phase, the mouth area is recognized and cropped from the original image.

**Figure 6: Examples for cropping the mouth**

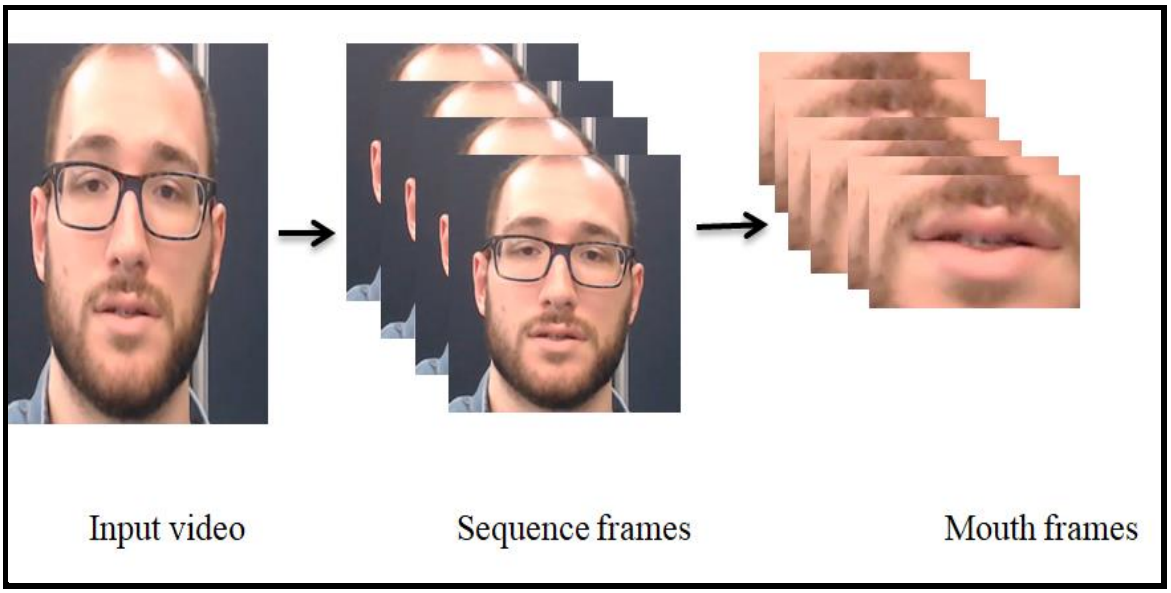


**Figure 7: Block diagram the preprocessing steps**

## Squeeze Net

This is a convolutional neural network that has been designed to be lightweight and compact, while still maintaining good performance on image classification tasks. The main aim of this network is to overcome the large size and high resource requirements of deep convolutional neural networks. The network was developed by scientist Basian Lapsica and his team at the University of California, Berkeley. It is used in a variety of applications, including image recognition, classification, and object detection. Due to its small size and strong performance, it can be deployed on mobile devices and in resource-limited systems[20].

One of the key innovations in SqueezeNet is its use of a method called "channel squeezing". This technique helps reduce the computational cost of the network without sacrificing its accuracy. It achieves this by decreasing the number of channels in the convolutional layers of the model. In addition to channel squeezing, SqueezeNet employs other methods such as fire modules and deep compression to further boost its overall efficiency [21].

**Squeeze Net layers**

This network consists of three layers [20]:

1- Convolution layers: That identify different features and patterns in images

2- Pooling layers: These layers reduce dimensions and remove some unimportant information.

3- Fire module layers: It classifies the images based on the information provided by the previous layers.

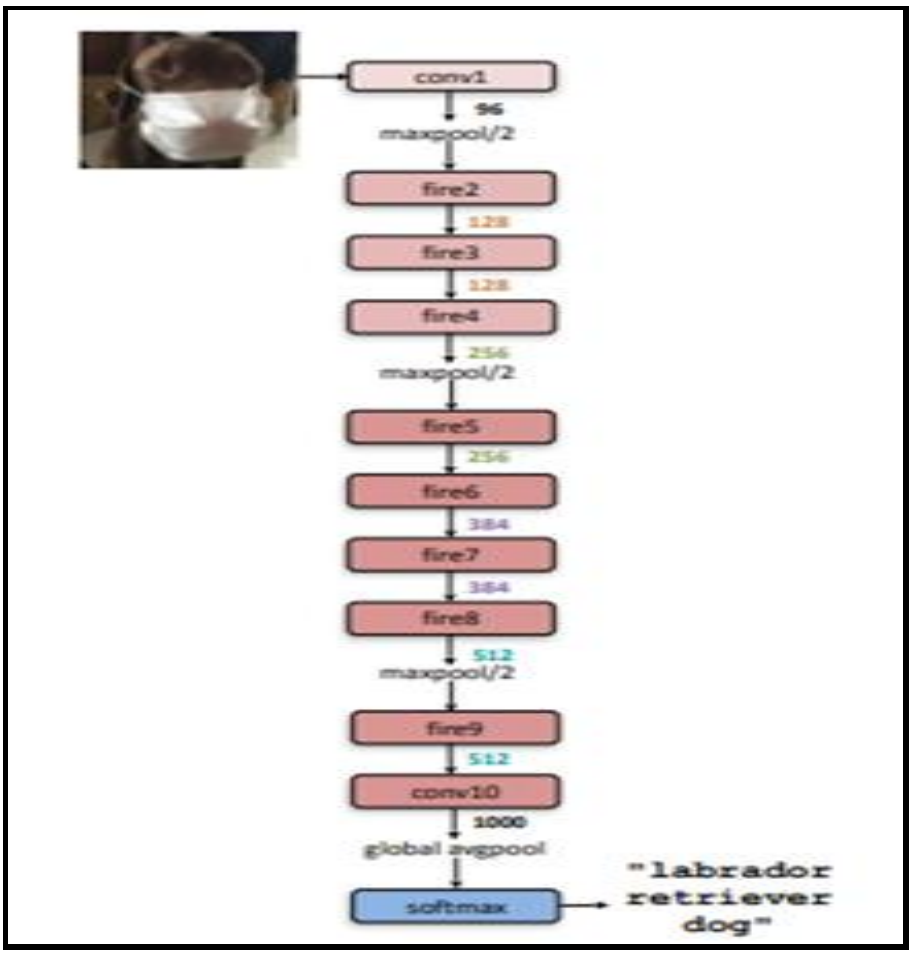


**Figure 8 : architecture of squeezeNet** [20]**.**

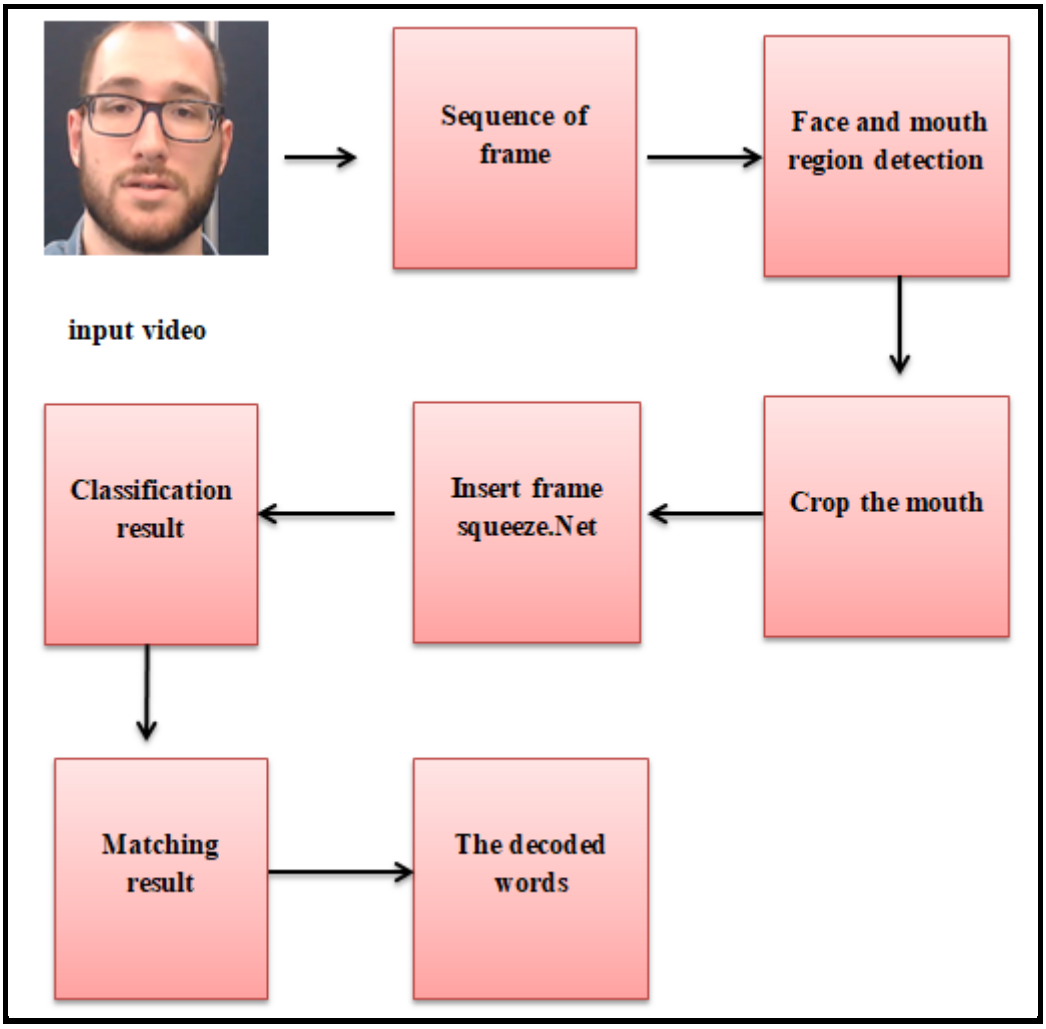


**Figure 9: The block diagram of the proposed system**

### The classification stage

The classification stage is the stage where the proposed system's performance is validated. In this stage, the mouth region frames are classified as either correct or incorrect readings, which is done by matching the frames from the training phase with the frames from the testing phase. In this stage, the video frames are inputted into the SqueezeNet network to obtain the classification results. It is important to note that the proper extraction and classification of features has a significant and obvious impact on the classification outcome. If the testing phase frames are correctly classified and match the training phase frames, This suggests that the network is functioning properly (correct reading) and the proposed system is working properly. In contrast, if the frames do not match, this indicates rate of errors in the network's functioning and, consequently, a malfunction in the proposed system.

## 4. Dataset Description

The database consists of 10 participants 6 men and 4 women, and every individual articulates the phrases 10 short phrases. The phrases are the same as those used in the OuluVS2 database: "Goodbye", "Excuse me", "Hello", "Nice to meet you", "See you", "How are you", "I'm sorry", "Have a nice time" , "Welcome" "Thank you", . And we observe 18000 frames and This was achieved by placing a

camera in front of the speakers of the phrases. The location where we recorded the videos is https://ibug-avs.eu/ .

## 5- Details of the Experiment

The number of video lips images used as input to the CNN , which were 32400 frames, within the convolutional neural network (Squeeze Net), where it traversed across the layers of the neural network. It the training operation and test process was 80% (24729 frames), and 20% (7680) frames respectively.

Training phase and test phase of Squeeze Net built upon MATLAB R2020 language to facilitate the implementation of CNN code. "A Lenovo computer with specifications that include an Intel® Core™ i7-10510U CPU running at 1.80GHz and 2.30GHz, 8.00 GB of RAM, Windows 10 Pro, and a 64-bit operating system with an x64-based processor."

### Squeeze Net effectiveness

With the Recommended approach, A precision rate of 90% was achieved Utilizing 6 participants, three males and three females. Squeeze Net attained success due to the challenges posed by the presence or absence of makeup for females, as well as the presence or absence of mustaches for males, and this database was sourced from the website https://ibug-avs.eu/.

## 6. Conclusion

The proposed system underwent multiple phases to achieve recognition of the pronunciation of phrases through lip movement. In the first stage, we use the (MTCNN) algorithm for face detection mouth region, Next isolate the area of interest (the mouth), save the lip frame in a temporary directory, and then input the frame into Squeeze Net to retrieve features and subsequently classify it . One of the benefits of the Recommended approach is that it provided accurate classification, even considering the variations in the structure and shape of the lips among different individuals.

## Reference

[1] V. C. Jadhav, R. M. Sonar, and S. D. Sancheti, “International Journal of Modern Trends in Engineering and Research,” 2016, *IJMTER*.

[2] A. N. C. Aarkar, “Roi Extraction and Feature Extraction for Lip Reading of,” *vol*, vol. 7, pp. 484–487.

[3] J. Huang, G. Potamianos, J. Connell, and C. Neti, "Audio-visual speech recognition using an infrared headset," *Speech Commun.*, vol. 44, no. 1–4, pp. 83–96, 2004.

[4] Y. Lan, R. W. Harvey, B.-J. Theobald, E.-J. Ong, and R. Bowden, "Comparing visual features for lipreading.," in *AVSP*, Citeseer, 2009, pp. 102–106.

[5] A. Garg, J. Noyola, and S. Bagadia, "Lip reading using CNN and LSTM," *Tech. report, Stanford Univ. CS231 n Proj. Rep.*, 2016.

[6] A. Mesbah, A. Berrahou, H. Hammouchi, H. Berbia, H. Qjidaa, and M. Daoudi, "Lip reading with Hahn convolutional neural networks," *Image Vis. Comput.*, vol. 88, pp. 76–83, 2019.

[7] A. K. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Trans. circuits Syst. video Technol.*, vol. 14, no. 1, pp. 4–20, 2004.

[8] Y. Lu and H. Li, "Automatic lip-reading system based on deep convolutional neural network and attention-based long short-term memory," *Appl. Sci.*, vol. 9, no. 8, 2019, doi: 10.3390/app9081599.

[9] Mr. Befkadu Belete Frew, "Audio-Visual Speech Recognition using LIP Movement for Amharic Language," *Int. J. Eng. Res.*, vol. V8, no. 08, pp. 594–604, 2019, doi: 10.17577/ijertv8is080217.

[10] M. Faisal and S. Manzoor, "Deep Learning for Lip Reading using Audio-Visual Information for Urdu Language," 2018.

[11] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 3444–3450, 2017, doi: 10.1109/CVPR.2017.367.

[12] G. Zhao, M. Barnard, and M. Pietikäinen, "Lipreading with local spatiotemporal descriptors," *IEEE Trans. Multimed.*, vol. 11, no. 7, pp. 1254–1265, 2009, doi: 10.1109/TMM.2009.2030637.

[13] A. A. Yasir, A. H. Hasan, and M. J. Hayawi, "Deep Learning Models for Classifying Driver Eyes: A Comparative Study.," *J. Coll. Educ.*, no. 4, 2021.

[14] S. Minaee, P. Luo, Z. Lin, and K. Bowyer, "Going Deeper Into Face Detection: A Survey," pp. 1–17, 2021.

[15] F. Zhao, J. Li, L. Zhang, Z. Li, and S.-G. Na, "Multi-view face recognition using deep neural networks," *Futur. Gener. Comput. Syst.*, vol. 111, pp. 375–380, 2020.

[16] M.-H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *IEEE*

*Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 1, pp. 34–58, 2002.

[17] S. Minaee, A. Abdolrashidi, H. Su, M. Bennamoun, and D. Zhang, "Biometrics recognition using deep learning: A survey," *Artif. Intell. Rev.*, vol. 56, no. 8, pp. 8647–8695, 2023.

[18] Z. Orman, A. Battal, and E. Kemer, "A study on face, eye detection and gaze estimation," *IJCSES*, vol. 2, no. 3, pp. 29–46, 2011.

[19] M. Wang and W. Deng, "Deep face recognition: A survey," *Neurocomputing*, vol. 429, no. July 2018, pp. 215–244, 2021, doi: 10.1016/j.neucom.2020.10.081.

[20] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size," no. April 2019, 2016.

[21] U. Kulkarni, S. M. Meena, S. V Gurlahosur, and G. Bhogar, "Quantization friendly mobilenet (qf-mobilenet) architecture for vision based applications on embedded platforms," *Neural Networks*, vol. 136, pp. 28–39, 2021.