Computer Systems
Science and Engineering

Tech Science Press

DOI: 10.32604/csse.2023.037113
Article

Visual Lip-Reading for Quranic Arabic Alphabets and Words Using Deep Learning

Nada Faisal Aljohani* **and Emad Sami Jaha**

Department of Computer Science, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, 21589, Saudi Arabia
*Corresponding Author: Nada Faisal Aljohani. Email: naljohani0084@stu.kau.edu.sa
Received: 24 October 2022; Accepted: 21 December 2022

Abstract: The continuing advances in deep learning have paved the way for several challenging ideas. One such idea is visual lip-reading, which has recently drawn many research interests. Lip-reading, often referred to as visual speech recognition, is the ability to understand and predict spoken speech based solely on lip movements without using sounds. Due to the lack of research studies on visual speech recognition for the Arabic language in general, and its absence in the Quranic research, this research aims to fill this gap. This paper introduces a new publicly available Arabic lip-reading dataset containing 10490 videos captured from multiple viewpoints and comprising data samples at the letter level (i.e., single letters (single alphabets) and Quranic disjoined letters) and in the word level based on the content and context of the book *Al-Qaida Al-Noorania*. This research uses visual speech recognition to recognize spoken Arabic letters (Arabic alphabets), Quranic disjoined letters, and Quranic words, mainly phonetic as they are recited in the Holy Quran according to Quranic study aid entitled Al-Qaida Al-Noorania. This study could further validate the correctness of pronunciation and, subsequently, assist people in correctly reciting Quran. Furthermore, a detailed description of the created dataset and its construction methodology is provided. This new dataset is used to train an effective pre-trained deep learning CNN model throughout transfer learning for lip-reading, achieving the accuracies of 83.3%, 80.5%, and 77.5% on words, disjoined letters, and single letters, respectively, where an extended analysis of the results is provided. Finally, the experimental outcomes, different research aspects, and dataset collection consistency and challenges are discussed and concluded with several new promising trends for future work.

Keywords: Visual speech recognition; lip-reading; deep learning; quranic Arabic dataset; Tajwid

1 Introduction

Human language is the most fundamental means of communication. Therefore, artificial intelligence algorithms have been developed to use spoken speech in different languages for automatic speech recognition. Speech recognition transforms spoken language into data bits that can be utilized for various applications. The field of automatic speech recognition has been gradually and increasingly a hot research area at different levels. It began with only audio speech recognition (ASR), then audio-visual speech recognition (AVSR), followed by continuous speech recognition (CSR), and finally, as the most challenging task among all its counterparts, performing only visual speech recognition (VSR). Lip-reading is a desired capability that can play an important role in human-computer interaction. In computer vision, recognizing of visual speech is a common complex research topic, as lip-reading can be affected by many factors such as lighting, age, makeup, viewpoint angle, and variation of lip shape.

On the other hand, a visual speech recognition system is not affected by surrounding noise because it does not require sounds for processing, as it is limited only to visual information processing. Lip-reading, or visual speech recognition, can be described as extracting spoken words based on visual information only through lips, tongue, and teeth movements made by the pronunciation of these spoken words. The significance of this research field appears in the possibility of applying the concepts of visual lip-reading to different languages for several applications, such as criminal scrutiny by surveillance cameras, educational evaluation, biometric authentication, and connecting with smart cars. Therefore, many research efforts are increasingly devoted to improving performance and increasing the accuracy of visual speech recognition.

The Arabic language is the language of the Holy Quran, meant to preserve and spread it in service to Islam, facilitating the recitation and understanding of the Quran. The Arabic language has elaborate and important origins, derivations, rules, amplitude, flexibility, and formation, which are not often found in other languages. Its letters are distinguished by their articulation (exits) and sound by harmony. Lip-reading may be one of the most important topics that can be effectively employed in human-computer interaction in multiple languages. However, the Arabic language still needs to be sufficiently studied by researchers, in contrast to other languages.

Recently, most research explorations related to the Holy Quran focused on only audio speech recognition; there has been almost no interest in employing the approaches of visual speech recognition or considerations of the potential improvement it would bring. The lack of Arabic visual speech recognition studies may affect progress and development in research related to the Holy Quran in visual speech recognition. The lack or unavailability of Arabic visual speech datasets is likely to decrease the number of studies in deep learning for the Arabic language. Addressing this problem will have practical benefits for studies based on Arabic and contribute to progress in Quranic research areas.

Deep learning is an emanative part of artificial intelligence and is designed in such a way as to simulate the strategies of the human brain. Its architecture contains several layers used for processing, similar to neurons. Deep learning algorithms have been found superior at most recognition tasks and are considered a vast improvement compared to traditional methods in various automatic recognition applications. Recently, some research studies have applied deep learning techniques to lip-reading and proved that they could be used successfully in VSR. Due to the tremendous success achieved by deep learning methods in many fields, we aim is to visually recognize Arabic speech using deep learning techniques based on the rules of the *Al*-Qaida *Al-Noorania book* [1] (***QNbook***), which is used in learning Quran recitation alongside **Tajwid**. **Tajwid** is a science that contains a set of rules that helps reciters of the Holy Quran pronounce the letters with the correct articulation (**Makharij**) and gives each letter its distinctive adjectives (**Sefat**). *QNbook* is considered one of the easiest and most

useful means of teaching the pronunciation of the Quranic words as the Prophet Muhammad; peace be upon him, recited them. By learning the smallest building block of the Quran, the letter, whoever masters it can read the Holy Quran by spelling without difficulty because of the author's precision and care in collecting it. The author started gradually with single letters, compound letters, disjoined letters, vowels, formation, and the provisions of Tajwid. Because of the importance of *QNbook*, the fruit of which is correct and eloquent pronunciation and a distinct ability to read Arabic in general and the Quran in particular, we worked on building a dataset that will benefit researchers and those interested in studies related to the Holy Quran in several areas. The most important of these is artificial intelligence to build promising studies in the future and spread its benefit. In this work, the main contributions are as follows:

- Building, to the best of our knowledge, the first new public dataset designed for audio-visual speech recognition purposes in the Arabic language according to the *Al-Qaida Al-Noorania book* (*QNbook*) from three different face/mouth angles (viewpoints). We called it Al-Qaida Al-Noorania Dataset (AQAND) for inducing a variety of further novel and effective research.
- Moving the field of speech recognition from audio to visual in studies related to the Holy Quran by training, validating, and testing the pre-trained convolutional neural network (CNN) model using our generated dataset, achieving an accuracy of 83%. This move may help to improve interactive recitation systems and tools as well as systems and tools for learning the recitation of the Holy Quran in the future.
- Recognizing, to the best of our knowledge, the single letters (Horof Alhejaā Almufradah) and disjoined letters (Horof Almuqataā) for the first time, using lip-reading and reliance on the classical Arabic language. At the word level, words from the Holy Quran were used rather than from various colloquial dialects used in many Arab nations. As far as we know, this has never been done before in visual speech recognition.

The paper is organized in the following manner: the related work is presented in Section 2; the AQAND's design, its pre-processing details, and the recommended classification model are all described in Section 3; Section 4 of the paper discusses the experiments and results; finally, Section 5 presents the conclusion and future work.

2 Related Work

In the last few years, many researchers have been working on VSR. Before the advent of deep learning, researchers used various methods and algorithms in machine learning for lip-reading. Where numerous research works in lip-reading were based on hand-engineered features that are usually modeled by hidden Markov model (HMM) based pipelines, as shown in [2,3]. There has been a significant advancement in lip-reading techniques during the past ten years. Many studies initially centered on 2D fully convolutional networks [4,5]. But as hardware improved, 3D convolutions over 2D convolutions [6–8] or recurrent neural networks [9,10] quickly became an option for more effective use of temporal information. This concept has developed into a specific architecture consisting of two parts: the front end and the back end. In this process, the final temporal information is summed using recurrent layers as a backend after the local lip movement information has been extracted as a frontend using a 3D + 2D convolutions backbone [11,12]. This proposed architecture [12] has been very effective. It has achieved a 17.5% recognizable improvement in accuracy in the lip-reading datasets LRW [13] and LRW-1000 [14], and even now, many state-of-the-art lip-reading solutions still use it as its meta-architecture [15–18]. Following the level of recognized speech, in this section, we present related work in the current research area by distributing it into four sub-sections:

2.1 Recognition of Arabic at the Letter Level

Arabic letter-level audio speech recognition researchers have achieved more than 99% accuracy [19,20]. Other than studies on visual speech recognition, most studies focus on numbers, words, and sentences rather than considering the alphabet, which forms the foundation of any language. Nevertheless, it is a must when learning any language to know the proper pronunciation of that language's alphabet using the correct articulations and exits. However, in [21], researchers classified the alphabet of the Arabic language into ten visemes (different letters having similar lips movements during pronunciation) and established its viseme mapping for four speakers. They based their study on an analysis of geometrical features of the face extracted from the front lip movement. As a result, they demonstrated that it is possible to recognize a vowel and determine whether it is a short vowel (i.e., fataha, damma, or kasra) or a long one. Also, the research of [22] presented an Arabic viseme system that classified the language into 11 visemes based on two methods: the statistical parameters for lip image and the geometrical parameters for the internal and external lip contour using multilayer perceptron (MLP) neural networks.

On the other hand, the work described in [23] is closely related to ours. The authors researched the mouth shapes of the alphabet during its pronunciation with Tajwid to find out the differences between its pronunciation from the geometry and sequence of movements of the lip and divided the 28 letters into five groups based on those movements. They introduced a lip tracking system to extract lip movement data from a single professional reciter and compare it with novice users' lip movements to verify their pronunciation's correctness through a graphical user interface (GUI) that they designed and programmed. Their study depends on calculating the displacement between the height and width of the lips for each video frame and plotting it in a graph using machine learning algorithms. As noted in the literature, there need to be more visual speech datasets for the Arabic alphabet that will open promising avenues of research in the future. In AQAND, we included Arabic and the 14 Quranic letters as novel content in letter-level recognition.

2.2 Recognition of Arabic at the Word Level

Most studies focused on lip-reading at the word level have achieved high results. In reference [3], the researchers proposed a novel approach that aims to detect Arabic consonant-vowel letters in a word using two HMM-based classifiers: one for the recognition of the "consonant part" and one for the "vowel part.". They tested their method on 20 words collected from 4 speakers. Their algorithm scored an accuracy of 81.7%. In addition to that study, [24] collected 1100 videos of 10 Arabic words from 22 speakers, calling it the Arabic visual speech dataset (AVSD). The authors cropped the mouth region manually from video frames, and the support vector machine (SVM) model was used to evaluate the AVSD with a 70% word recognition rate (WRR). Also, [25] presented the read my lips (RML) system, an Arabic word lip-reading system. The authors collected dataset videos of 10 commonly used Arabic words from 73 speakers, with RGB and grayscale versions of each. They trained and tested the dataset using three different deep-learning models, as listed in Table 2. The RGB version of the dataset obtained higher accuracy than the grayscale. They also suggested a voting model for the three models in the RML system to improve the overall accuracy, as they succeeded and achieved an accuracy of 82.84%, which is 3.64% higher than the highest accuracy obtained for one of the three models. This accuracy was the highest Arabic word prediction accuracy in any related work. However, in our Arabic Quranic word dataset, we scored a higher accuracy by 0.5%.

2.3 Recognition of Arabic at the Sentence Level

At this level, only a few studies have been conducted on lip-reading. Over the last two decades, researchers have proposed for the first time a novel Arabic lip-reading system by combining the hypercolumn model (HCM) with the HMM [2]. They used HCM to extract the relevant features and HMM for feature sequence recognition. For testing their proposed system, they used nine sentences uttered by 9 Arabic speakers and achieved 62.9% accuracy. Recently, researchers continued to improve the recognition process and performance at this level, as shown in [26]; they presented an Arabic dataset of sentences and numbers for visual speech recognition purposes, which contains 960 sentence videos and 2400 number videos from 24 speakers. They used concatenated frame images (CFIs) as pre-processing for their dataset, resulting in one single image containing utterance sequences, which are fed into their proposed model: visual geometry group (VGG-19) network with batch normalization for feature extraction and classification. They excelled, achieving a competitive accuracy of 94% for numbers prediction and 97% for sentence prediction.

2.4 Holy Quran-Related Studies in Speech Recognition

In recent decades, researchers have presented many studies related to the Holy Quran, which are concerned with helping to learn and recite the Holy Quran correctly [27–29]. The researchers were also interested in several studies in the field of the seven readings recognition (**Qiraat**) and distinguishing it, such as in [27,30]. However, while these studies have shown remarkable performance in audio speech recognition, they have yet to be involved in visual speech recognition (VSR) in Quranic-related work areas. In terms of the **Qiraat**, we take into consideration that there is some confusion between two sciences (**Qiraat science** and **Tajwid science**) which must be clarified. Qiraat science is focused on how some words in verses of the Quran should be pronounced, while Tajwid science is more focused on the letters, their exits, and the attributes they carry. This means that the issue of "exits of letters" is therefore included under Tajwid science. Additionally, because the issue of exits deals with each letter separately, the variations between Qiraat do not affect this issue. In our work, we left the reciters the freedom of choice in pronouncing the Quranic words during data collection. However, by including the different Qiraat competitive accuracy was achieved.

Eventually, we conclude that deep learning networks have proved to be more efficient compared to other approaches. There is also a need to design and build open-source large-scale Arabic datasets of visual speech recognition, enabling more research efforts on this rich language. So, this research is an invitation to all researchers interested in the Holy Quran and Arabic visual speech recognition to continue and advance our work. Visual speech recognition datasets are summarized in Table 1. Table 2 summarizes all Arabic-related work described above.

Table 1: Arabic and other languages lip-reading datasets statistics

Language	Dataset	Dataset contents	Number of speakers	Recording environment	Resolution	Open source
Arabic	[2] 2004	9 Arabic sentences	9	Lab	160 × 120 pix	No

(Continued)

Table 1: Continued

Language	Dataset	Dataset contents	Number of speakers	Recording environment	Resolution	Open source
	AVSD [24] 2019	1100 video samples of 10 daily communication Arabic words	22 (8 males & 14 females)		1920 × 1080 pix	No
	[25] 2022	1051 video samples of 10 common Arabic words	73 (40 males & 33 females)		66 × 100 pix	Yes
	[26] 2022	3360 video samples of 10 Arabic digits & 4 Arabic sentences	24 (14 males & 10 females)		1920 × 1080 pix	Yes
	AQAND (ours)	10490 video samples of 29 Arabic alphabets & 14 Quranic letters & 10 Quranic words	22 (18 males & 4 females)		1920 × 1080 pix	Yes
English	LRW [13] 2016	500 words	+1K	TV	256 × 256 pix	Yes
Mandarin	LRW-1000 [14] 2019	1000 words	+2K	TV	Naturally distributed	Yes
Russian	LRWR [15] 2021	235 words	135	YouTube	1920 × 1080 pix	Yes

Table 2: Recent Arabic VSR-related work

Reference	Pre-processing Techniques		Features extraction	Classifier	Level recognition	Results and accuracy
	Face detection	Lip localization				
[2] 2004	N/A	N/A	Hypercolumn model (HCM)	HMM, with five states	9 Arabic sentences	Accuracy of 62.9%

(Continued)

Table 2: Continued

Reference	Pre-processing Techniques		Features extraction	Classifier	Level recognition	Results and accuracy
	Face detection	Lip localization				
[3] 2011[1]	N/A	N/A	Using low-level statistical methods, calculated the pixels of the ROI for extraction of geometrical features in 4 points on lips (W, H, A, D).	HMM with three states	20 Arabic words	Accuracy of 81.7%
[24] 2019	N/A	Manually cropped for ROI	DCT	SVM	10 Arabic words	Word recognition rate of 70%
[25] 2022	Python Dlib library	Dlib facial landmark points and generate two versions: RGB & grayscale	Three models: CNN & TD + LSTM & TD + BiLSTM	Softmax layer	10 Arabic words	Accuracy of: RGB in CNN (79.2%), grayscale in CNN (76.6%), RGB in TD + LSTM (70.1%), grayscale in TD + LSTM (67.5%), RGB in TD + BiLSTM (74.1%), grayscale in TD + BiLSTM (70.1%), and RGB in a voting model (82.8%)
[26] 2022	OpenCV Library using Dlib toolkit to detect facial landmarks	Use facial landmarks to locate key points of mouth and generate CFI	VGG-19 with batch normalization	Softmax layer	10 Arabic digits and 4 Arabic sentences	Accuracy of digits is 94%, in sentences is 97%, in digits & sentences is 93%

Note: [1]Speaker-independent

3 The Al-Qaida Al-Noorania Dataset (AQAND)

Most lip-reading systems involve two phases (analyzing visual information in the input image and transforming this information into corresponding words or sentences) and three parts (lip detection and localization, lip feature extraction, and visual speech recognition). Many studies have given end-to-end models driven by deep neural networks (DNN) to hasten the development of deep learning technologies. These models have shown promising performance outcomes when compared to approaches based on traditional neural networks. In this research, we aim to train and evaluate an

effective model for lip-reading through three main phases: AQAND design and collection, AQAND preparation and splitting, and implementation of the classification model.

3.1 AQAND Design

In this research investigation, we built our dataset since no datasets were available in our research scope nor suitable to achieve our objectives. Therefore, we introduce a new and publicly available lip-reading dataset, which, to the best of our knowledge, may be the first dataset on *QNbook*. As such, the AQAND is designed and created to be used for training and testing a model's capabilities initially in lip-reading (recognition of the lip movements) and further, in recognizing the correct and accurate exits of the letters during their pronunciation by following the intonation (Tajwid) rules for Quran recitation. The AQAND consists of around 16 h of RGB videos with a resolution of 1920 × 1080 pixels and 30 frames per second, resulting in a total of 10490 RGB video samples, each ranging from two to ten seconds long. The total size of AQAND is approximately 60 GB.

3.2 AQAND Collection

The collected videos of the dataset were sorted into three main categories: Horof Alhejaā Almufradah (29 single letters), Horof Almuqataā (14 disjoined letters), and Quranic words (10 words). Each video was recorded in an indoor environment from three different angles (0°, 30°, and 90°) simultaneously using three digital cameras, with each camera placed within 90 cm of the subject. All videos were obtained from 22 reciters (contributors), 16 men and four women from the ages of 20 to 59. Each reciter repeated the same video recording scenario (for all three categories) in three different sessions. In addition, there were several individual differences between reciters, including geometric features of the lip, beard, mustache, movements, and shape of mouth, teeth, and the alveolar ridge. There were also observable variations that occurred between the same repeated samples over the three recording sessions from the same reciters, which helped in producing unbiased data sampling and can enable data augmentation for deep learning.

The proposed setup and recoding scenario were to allow the following of Tajwid rules based on the *QNbook* method, where the reciters were asked to read letters and words correctly, clearly, and accurately and wait for five seconds (a silence between each letter/word) before moving to the next item. After taking the videos, we entirely reviewed each video to verify their authenticity and to make sure that our instructions were correctly followed. As such, the videos were collected for the three main categories (*QNbook* lessons as shown in Fig. 1) and included in the AQAND as follows:

3.2.1 Arabic Single Alphabets Dataset

The first and core lesson in *QNbook* is Horof Alhejaā Almufradah (29 single letters), as in Fig. 1a, where their pronunciation is clarified with English letters in Table 3. A reciter reads the single letters as written in red above each letter on the *QNbook* page, shown in Fig. 1a. This means they utter it independently, as the letter's conventional full name and not the letter's isolated sound (e.g., for 'ا,' saying "alif" and not "a'a") as per the *QNbook* rules. This dataset consists of 5742 videos.

(a) (b) (c)

Figure 1: The three main categories contained in the AQAND as presented in *QNbook*: (a) Single alphabets (Horof Alhejaā Almufradah), (b) Disjoined letters (Horof Almuqataā), and (c) The selected ten Quranic words that are highlighted in red boxes

Table 3: The Arabic letters in isolated forms, their corresponding forms in English, the letters in Arabic script, and the pronunciation of single Arabic letters in English. (Note that it is difficult to write the exact spelling of some Arabic letters as there are no similar phonemes to them in the English language)

Alphabets sequence	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Isolated forms in Arabic	أ	ب	ت	ث	ج	ح	خ	د	ذ	ر	ز	س	ش	ص	ض
Isolated form in English	a	b	t	th	j	h	kh	d	dh	r	z	s	sh	s	dh
Letters with Arabic script	ألف	با	تا	ثا	جيم	حا	خا	دال	ذال	را	زا	سين	شين	صاد	ضاد
Pronunciation in English	Alif	Ba	Ta	Tha	Jeem	Ha	Kha	Dal	Dhal	Ra	Za	Sien	Sheen	Sad	dhad
Alphabets sequence	16	17	18	19	20	21	22	23	24	25	26	27	28	29	
Isolated forms in Arabic	ط	ظ	ع	غ	ف	ق	ك	ل	م	ن	و	هـ	ي	ء	
Isolated form in English	t	dh	a	gh	f	q	k	l	m	n	w	h	y		
Letters with Arabic script	طا	ظا	عين	غين	فا	قاف	كاف	لام	ميم	نون	واو	ها	يا	همزة	
Pronunciation in English	Ta	Dha	Aieen	Ghein	Fa	Qaf	Kaf	Lam	Meem	Noon	Wow	Ha	Ya	Hamzah	

3.2.2 Disjoined Letters Dataset

We also shed light on the third lesson from the lessons of *QNbook*, which is Horof Almuqatah (disjoined letters or Quranic letters), as shown in Fig. 1b. They are the fourteen combinations of disjoined letters stated in the Quran at the beginning of some surahs, and they are read as individual letters with no silence between them. For example, (الٓمٓ) is read connectedly as 'alif, lam, meem.'. The

reciters were asked to correctly perform the extension of letters' vowel segments if they are marked above by the (~) sign, according to the correct length (or duration, known in Tajwid science as the number of moves; in this case, six moves) for the marked letter. Table 4 clarifies the extended letter pronunciation by six moves in English by repeating the extended vowel letter, such as using 'Qaaaaaaf' for the extended letter (ﻗّ) instead of 'Qaf' as used for the unmarked standard letter (ق), shown in Table 3. This dataset consists of 2772 videos.

Table 4: Extend letter pronunciation by six moves of Arabic disjoined letters written in English. (Note that it is difficult to write the exact spelling of some Arabic letters because there are no similar phonemes to them in the English language)

Letters sequence	1	2	3	4	5	6	7
Disjoined letter	الٓمٓ	الٓمٓصٓ	الٓر	الٓمٓر	كٓهٰيٰعٓصٓ	طٰه	طٰسٓمٓ
Pronunciation in English	Alif laaaaaam meeeeeem	Alif laaaaaam meeeeeem saaaaaad	Alif laaaaaam ra	Alif laaaaaam meeeeeem ra	Kaaaaaaf ha ya aieeeeen saaaaaad	Ta ha	Ta sieeeeen meeeeeem
Letters sequence	8	9	10	11	12	13	14
Disjoined letters	طٰسٓمٓ	يٰسٓ	صٓ	حٰمٓ	حٰمٓعٓسٓقٓ	قٓ	نٓ
Pronunciation in English	Ta sieeeeen	Ya sieeeeen	Saaaaaad	Ha meeeeeem	Ha meeeeeem aieeeeen sieeeeen qaaaaaaf	Qaaaaaaf	Noooooon

3.2.3 Quranic Words Dataset

In this part, we have collected data for ten Quranic words from the sixth and ninth lessons of *the QNbook*. They were randomly chosen from among several Quranic words. A reciter was asked to read the ten Quranic words correctly with diacritics (signs above or under letters, which affect letter pronunciation as short vowels), as shown in the red boxes in Fig. 1c. Table 5 shows the pronunciation and meaning of the ten chosen Arabic Quranic words as written in English. This dataset consists of 1980 videos.

3.3 Pre-Processing the AQAND

In lip-reading, pre-processing is a necessary step, primarily affecting the validity and accuracy of recognition tasks. In this section, we present the conducted pre-processing steps. First, during pronunciation, each letter or word may have a different length. Therefore, the length of each video sample in the dataset is reformed to exactly 60, 80, or 300 frames, and each video is ensured to accurately contain the complete visual representation of the target letter or word. The number of frames is selected to range from 2 to 10 seconds because most reciters are observed to spend time within this range to complete an utterance of one letter or word. In a few instances, when letter/word utterance is performed faster than this predetermined period, the whole video is concatenated with additional black frames by using the zeros function to compensate for the missing frames. Thus, we get inputs that have a fixed sequence of frames of 60, 80, and 300 and durations of 2, 2.5, and 10 s for single letters, Quranic words, and disjoined letters, respectively.

Table 5: Pronunciation and meaning of the ten Quranic words in the dataset

#	Pronunciation of word		Meaning in English	#	Pronunciation of word		Meaning in English
	in Arabic	in English			in Arabic	in English	
1	عَذَابٌ	Adhaab	Punishment	6	سِرَاجًا	Seraja	Lamp
2	وَيْلٌ	Whyl	Woe	7	مَلِكِ	Maleke	Sovereign
3	حَسَدَ	Hasad	Envies	8	خُلِقَ	Kholeeqa	Created
4	عِنَبًا	Enaba	Grape	9	أَحَدٌ	Ahad	One and only
5	نُوحٍ	Lawoh	Slate	10	كُفُوًا	Kofua	equivalent

Second, if the data samples include unneeded details, this may adversely affect the lip-reading task. To combat this, the AQAND is designed to present a region of interest (ROI) in videos, such that the data sample is pre-processed using the Haar-cascade of OpenCV to detect and extract the ROI (in this case, the mouth region) from each video, as shown in Fig. 2. In visual speech recognition, the pre-processing works to accurately look for the position of the lip needed to adjust for any challenges, such as variations in positions of the face and mouth, variation in illumination, and image shadow, all of which can affect the quality of lip video frames. Therefore, a lip-reading system relying on accurate localization of the lip is more likely to achieve higher accuracy in visual speech recognition. A few samples of video frames are presented in Fig. 3, which display the sequence of visemes of the word 'Lawoh' for the same reciter from three angles (0°, 30°, and 90°).

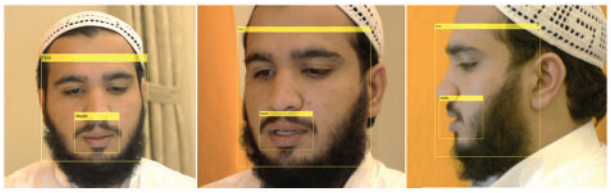


Figure 2: Face and mouth detection in video frames captured from three viewpoints (0°, 30°, and 90°)

Figure 3: The sequence frames of pronunciation of the word 'Lawoh' from 0° (frontend), 30°, and 90° (profile), up-to-down, respectively, where a voice letter is written under each corresponding frame

Third, to decrease the size of the dataset while maintaining reliable data quality, each frame is resized to 88 × 88 pixels, which is the input size used to train our model. Then, the resized frame segments are converted to grayscale, and their pixel values are normalized to a range between 0 and 1.

A cross-validation technique with a ratio of 70:30 had used to split normalized data samples into three sets for training, validation, and testing. Note that this work is conducted based on speaker-independent experiments (i.e., it seeks to identify anyone's lip movements for certain spoken letters/words, regardless of the speaker). All reciters are exclusively distributed into the three groups, everyone in only one of the groups, based on the variation between reciters in the same group in terms of skin color, gender, and age. The validation set is used to fine-tune the model's weights and biases, enhance performance, and prevent overfitting, while the training set is used to train and fit the deep neural network model. Unseen data (i.e., unseen reciters) is utilized in the testing set to assess the model's ability to recognize and classify spoken letters or words. The training set consists of 7630 videos, the validation set constitutes 1430 videos, and the testing set comprises 1430 videos. The size of the training and validation sets is twice doubled using data augmentation techniques, including horizontal flipping and an affine transformation (i.e., training and validation sets end with 22890 and 4290 videos, respectively). Finally, the pre-processed and labeled dataset is fed to the network model as an input with dimensions $88 \times 88 \times 1$, as will be further discussed in Subsection 3.4. Fig. 4 summarizes all AQAND pre-processing steps described above. The deep learning model in use automatically extracts features from image frames that include lip movements and then uses those features to classify the input video of spoken letters or words. Thus, the model can be utilized to test and classify an unseen input video of a completely unseen reciter. The proposed AQAND specifications are summarized in Table 6.

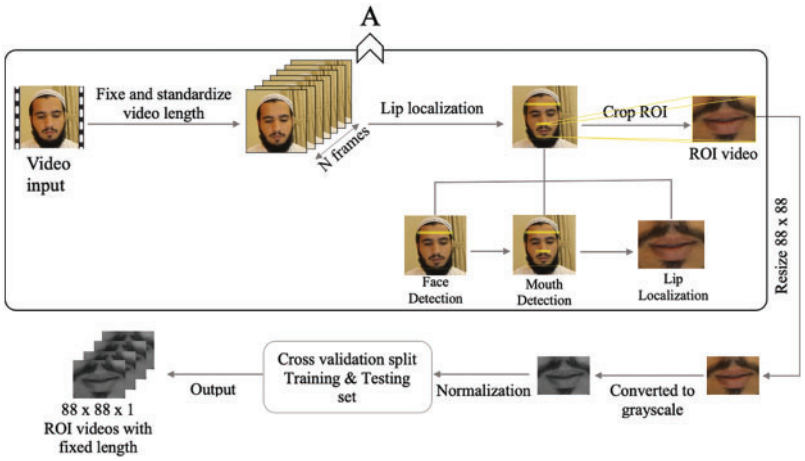


Figure 4: Flow chart of whole pre-processing. *N* refers to the fixed number of video frames. Section A is applied only one time, while the remaining steps include data pre-processing that is related to classification and feature extraction for the model and can be replaceable

Table 6: An overview of the AQAND dataset specifications

Specification aspect	Factor of specification	Specifications
Language	Language	Arabic
	Type of utterance	Single letters & disjoined letters & Quranic words
	Number of utterances	53 utterance classes

(Continued)

Table 6: Continued

Specification aspect	Factor of specification	Specifications
Reciters (contributors)	Number of reciters	22 reciters
	Gender	18 males and four females
	Number of repetition of utterances	Three times
	Face view angle of the reciter (viewpoint)	0° (frontend), 30°, and 90° (profile)
	Speaker-independent (acquisition/usage)	Yes
Technicality	Cameras	Three digital cameras
	Output data	Videos with MOV format
	Resolution	Full HD (1920 × 1080 Pixel)
	Frame rate	30 fps
	Method for controlling vibration	Using a camera tripod stand
	Recording environment	Controlled lab with good illumination and plain background

3.4 Classification Model

The transfer learning technique was used to investigate various efficient lip-reading deep learning models and examine their classification capabilities on our video data from AQAND, as shown in Subsections 3.1 and 3.2, to enforce a reliable deep learning model capable of achieving a high recognition accuracy for visual speech recognition. In this work, based on transfer learning, we apply AQAND, a modified state-of-the-art method [12], represented in Fig. 5, and evaluate the performance efficacy of the nascent model, then give a thorough analysis of the findings and insights for future research. The classification model is given an input video after pre-processing and applying data augmentation techniques where "HF + AT" in Fig. 5 means that horizontal flip and affine transformation techniques are applied. Then, we fed it into the residual network (ResNet-18), modifying 2D convolution to 3D. The output features from the global average pooling are then given to the backend network. The total number of letter or word classes is the output dimension of the final fully connected layer.

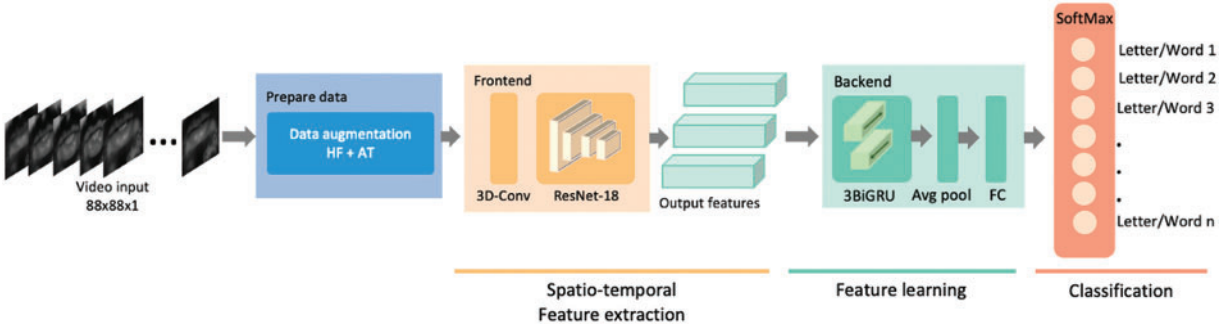


Figure 5: The overview of workflow a classification model architecture in our task

4 Experiments and Results

4.1 Experimental Setup and Initialization

The experiments are carried out utilizing the Google Colab environment with a large amount of RAM (51 GB) and a single GPU. The Google Drive is connected and accessed by the Colab environment, where all data and necessary files are stored. The Pytorch library is used for coding the proposed classification model outlined in Section 3. The Cross-Entropy (CE) loss function for optimization and the Adam optimizer with an initial learning rate of 7e-4 is used and combined with cosine learning rate scheduling and weight decay of 1e-4 with a batch size of 32. The cosine learning rate scheduling training trick is used in the same manner as in [12]. Every time the validation error reaches a plateau throughout three consecutive epochs, the learning rate will be reduced by a factor of two when validating the model at the end of each epoch. This is to prevent any abrupt reduction of the learning rate. The learning rate η at an epoch τ in the cosine setting is determined by Eq. (1) as follows:

$$\eta_{\tau} = \frac{1}{2}(1 + cos\left(\frac{\tau\pi}{T}\right))\eta \tag{1}$$

where η represents the initial learning rate, and T stands for the overall number of epochs, which in our experiments is 100.

4.2 Experimental Dataset Statistics

In this paper, all conducted lip-reading experiments are performed only on the zero-angle video samples (i.e., frontend) of AQAND, with all categories (classes) of the dataset (single letters, disjoined letters, and Quranic words), consisting of a total of 53 letters and words classes, obtained from 22 reciters repeated three times, resulting in 3498 total samples, which we split into 2544 training samples, 477 validation samples, and 477 testing samples. Table 7 shows the number of samples distributed in each category per split set. Each category of data is trained individually.

Table 7: Number of samples per split set of zero-angle video data

Dataset split	Splitting ratio	Number of reciters	Main categories in the dataset		
			Single alphabets	Disjoined letters	Quranic words
Training set	70%	16	1392	672	480
Validation set	15%	3	261	126	90
Testing set	15%	3	261	126	90

4.3 Experimental Analysis

4.3.1 Arabic Speech Production

Each letter in every language has a specific sound. Each sound is produced by a speaker from a specific place of the mouth using a precise utterance mechanism. Twenty-nine letters make up the Arabic alphabet, as per *QNbook*. The Arabic language can be distinguished from other languages in that each letter has a specific phoneme, and some of these letters are unique and cannot be found in other languages, such as the letter 'Dhad,' which in Arabic is written as 'ض.'. Fig. 6 explains all 29 Arabic letters, and from which place in the mouth they are each produced (also described as the

phonic production/exit place of a spoken letter). It also shows that Arabic letters are produced from 13 different places with specific phonemes and lip movements for each. All letters are arranged in groups of different colors according to their phonic production place in the mouth, starting with the letters issued from the lips and ending with the letters issued from the back of the throat (glottal).

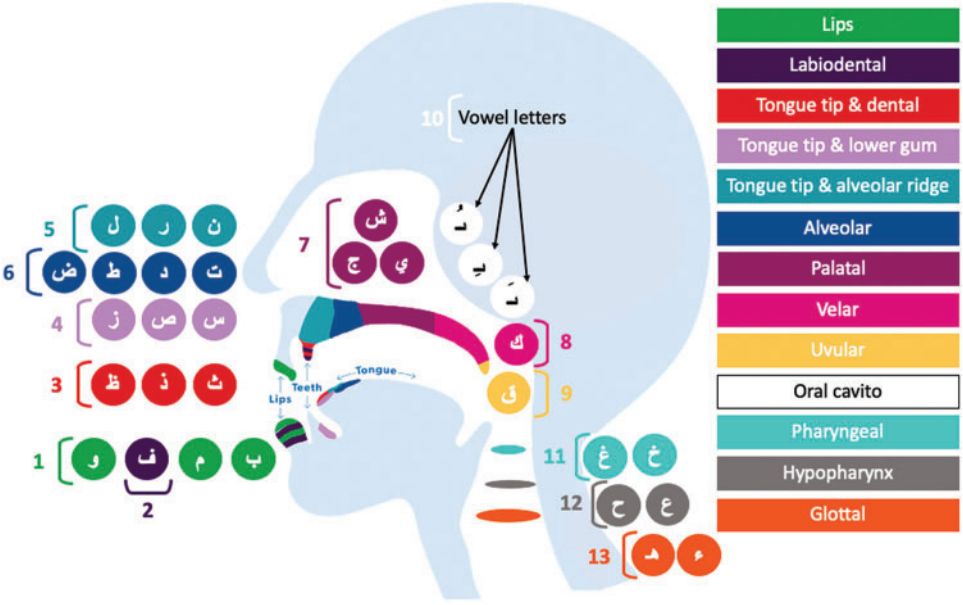


Figure 6: An illustration of Arabic letters production/exit places of the mouth during each letter utterance

The Arabic-based visual speech recognition task may face many challenges regarding the letters in groups 11, 12, and 13, as shown in Fig. 6. These letters are distinguished in the Arabic language and cannot be found in most other languages. Recognizing these letters only visually is considered challenging as they are produced from the back of the throat with no lip movement. This difficulty in recognizing these letters is also considered a major problem for deaf persons because they cannot notice slight changes in the movement of the lips, so they may not be able to imitate and pronounce these letters. As such, in visual speech recognition, we must pay attention to the concept of "visemes". Visemes are letters or words that are similar in the visual mouth movement and shape during their pronunciation but are different in sound and meaning. Table 8 summarizes all visemes found in the Arabic alphabet. This includes the group of letters 'Tha' (ثا), 'Dha' (ظا), and 'Dha' (ذا), which are three completely different letters in sound, but their visual movements look very similar on the speaker's lips, as shown for viseme number 3 in Table 8.

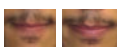
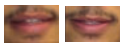
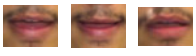
Focusing on the pronunciation style of the Arabic alphabet was the first step in the analysis, and we observed that it relied on two factors:

- **The ARTICULATIONs** are the places of the letter originates (area of letters exit from the mouth), which are identified and summarized per similar shapes of the mouth and are called visemes. Some Arabic letters share the same *articulations* and, thus, mimic the same movements of the lips (i.e., the same viseme), which creates a challenge to recognize and classify them. Note that the entire alphabet is categorized into 12 visemes, as shown in Table 8.

- **The HOWs**, which are the styles or the manners that accompany each letter when pronouncing it, such as 'whispering' ('Hams'), 'loudness' ('Jahr'), 'intensity' ('Shedah'), and 'looseness' ('Rakhawah'), etc. They vary from strong to weak, and the letter's strength or weakness is determined by the number of *hows* it carries and whether strong or weak.

Since our work is isolated from audio speech recognition and dependent solely on visual speech recognition, we divided the single letters based on their exits, strength, and weakness in pronunciation into three groups, so the letters with similar visemes are separated into two different groups. The *hows* have a key role in distinguishing between similar visemes; this is their primary function. If we had dealt with audio speech recognition along with visual speech recognition, the audio information of *hows* would have had an additional role in discriminating between the letters with visually similar visemes.

Table 8: The twelve visemes of Arabic isolated letters, illustrating the visual similarity between letters in the same group that have the same exits, resulting in one single inferred viseme for them

Viseme number	1	2	3	4	5	6
Isolated Arabic letter	[ب ، م]	[و]	[ث، ذ، ظ]	[ط، ض]	[ر، ل، ن]	[خ، غ، ق]
English Pronunciation	[ma, ba]	[wa]	[dha,dha,tha]	[dha, ta]	[na, la,ra]	[qa, gha,kha]
The similarity of each letter		unique viseme				
Inferred viseme						
Viseme number	7	8	9	10	11	12
Isolated Arabic letter	[ت، د]	[ف]	[س، ز، ص]	[ح،ك،ع]	[أ، هـ، ء]	[ش، ج، ي]
English Pronunciation	[da,ta]	[fa]	[ša,za,sa]	[aa,ka,ha]	[āa,ha,aa]	[ya,ja,sha]
The similarity of each letter		unique viseme				
Inferred viseme						

4.4 Experimental Results and Discussions

The confusion matrix of the classification model using the testing set for the Quranic words category is shown in Fig. 7. The number of samples in the testing set for each class is represented by the sum of the numbers in each row, which is constant (i.e., nine testing samples from 3 reciters). Since the mouth movements of the letters in words W3 ('Hasad,' حسد), W6 ('Seraja,' سراجا), and W9 ('Ahad,' أحد) are so similar, it is apparent that W3 is the most confusable and difficult to predict accurately. In the sequence of frames for the words W3, W6, and W9, respectively, as shown in Fig. 8, the visemes are clear to see. Another significant observation is that the words W2 ('Lawoh,' لوح), W8 ('Kholeeqa,' خلق), and W10 ('Kofua,' كفوا) are very similar in their use of an 'O' mouth shape, which creates a challenge in differentiating and predicting them accurately. The loss and overall classification accuracy of the Quranic words in the testing set resulted in 83.33% accuracy, as shown in Table 9, beside each class's precision and recall statistics.

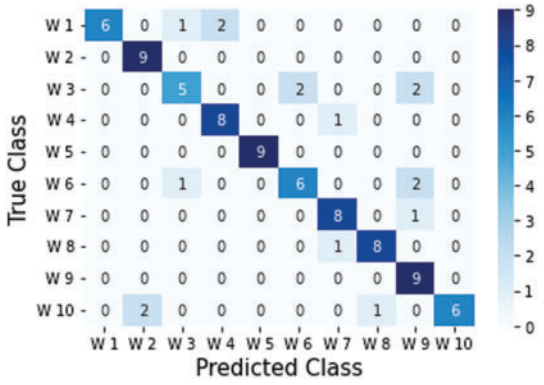


Figure 7: The confusion matrix for Quranic words classification

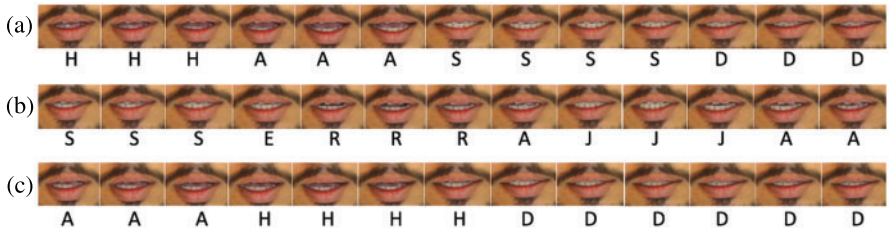


Figure 8: Pronunciation sequence frames for three words: (a) W3 ('Hasad,' حسد), (b) W6 ('Seraja,' سراجا), and (c) W9 ('Ahad,' احد). Each letter (approximate corresponding voice) is written under each frame

Table 9: The precision, recall, loss, and testing accuracy for Quranic words

Class	Adhaab (W1)	Whyl (W2)	Hasad (W3)	Enaba (W4)	Lawoh (W5)	Seraja (W6)	Maleke (W7)	Kholeeqa (W8)	Ahad (W9)	Kofua (W10)
Precision	100	81.8	71.4	80	100	75	80	88.9	64.3	100
Recall	66.7	100	55.6	88.9	100	66.7	88.9	88.9	100	66.7
Overall accuracy 83.33%										
Loss 1.07										

The confusion matrix of the disjoined letters category for the testing set is shown in Fig. 9. We observed that the letters DL8 ('TaSieen,' طس) and DL9 ('YaSieen,' يس) caused the most mutual confusion in classification due to visemes used in the second part of the word. The loss and overall classification accuracy of the disjoined letters in the testing set to result in an accuracy of 80.47% with corresponding precision and recall results for each class are shown in Table 10.

Given the single letters splitting method outlined in Sub-section 4.3, the confusion matrices for testing the three groups (G1, G2, and G3) of single letters are shown in Fig. 10. It appears that single Arabic letters are the most difficult classification task compared with the other two categories. This is due to several letters possessing the same exits, which results in sharing similar visemes. Furthermore, such spoken letters are short, and some of them have no apparent movement of the lips because the spoken letters are produced from the back of the throat (Horof Halaqia). We obtained accuracies of 72.66%, 70.31%, and 77.48% for G1, G2, and G3, respectively, despite the challenging factors affecting the performance of the classification model in this task. The loss and overall classification accuracy for each group of the single letters test set and precision and recall statistics are shown in Tables 11, 12, and 13 for G1, G2, and G3, respectively.

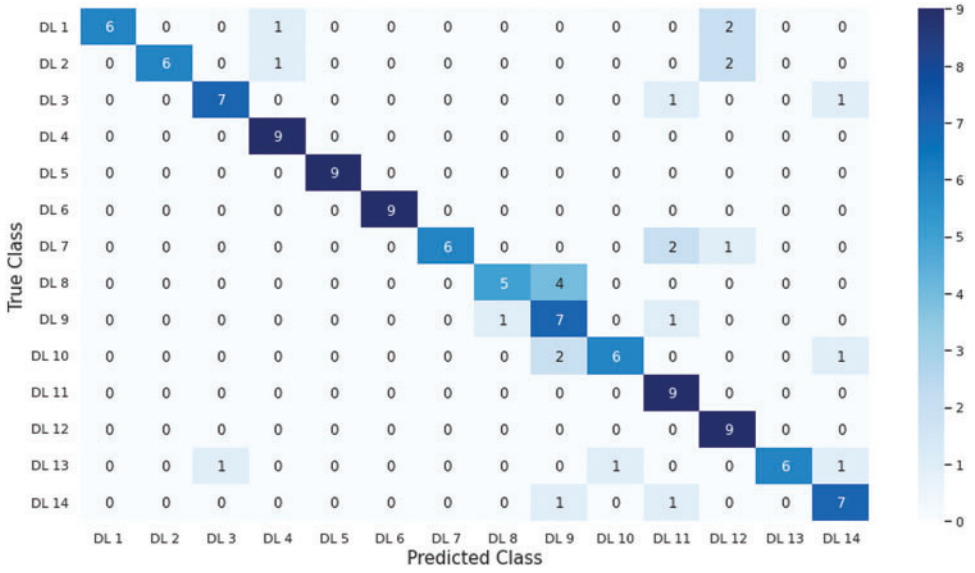
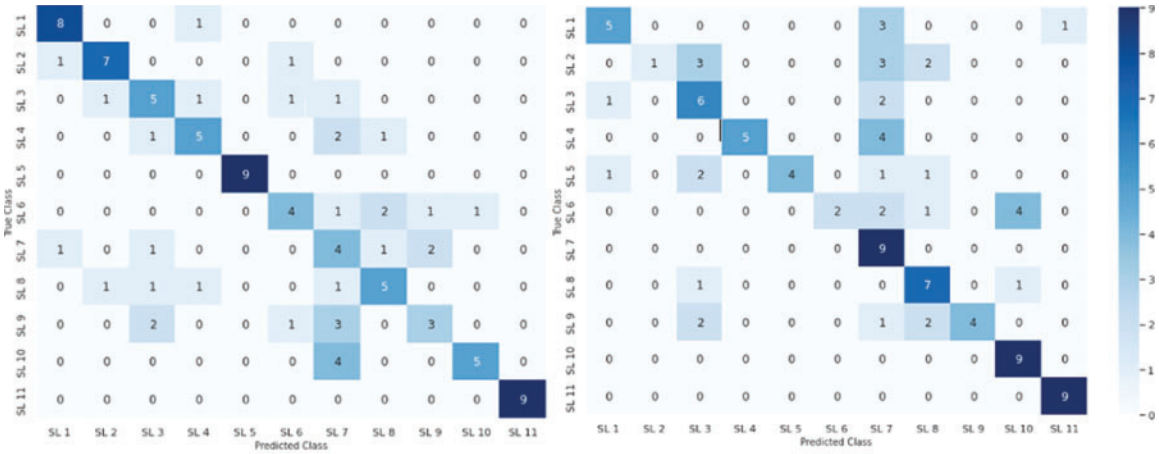


Figure 9: The confusion matrix for disjoined letters classification

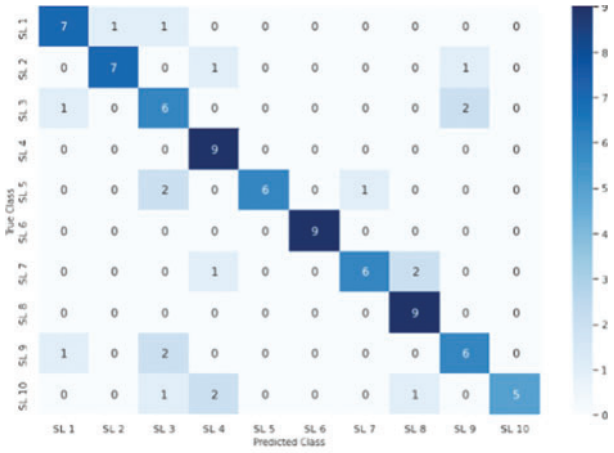
Table 10: The precision, recall, loss, and overall accuracy of disjoined letters

Class	DL1	DL2	DL3	DL4	DL5	DL6	DL7	DL8	DL9	DL10	DL11	DL12	DL13	DL14
Precision	100	100	87.5	81.8	100	100	100	83.3	50	85.7	64.3	64.3	100	70
Recall	66.7	66.7	77.8	100	100	100	66.7	55.6	77.8	66.7	100	100	66.7	77.8
Overall accuracy 80.47%														
Loss 1.5														



(a) (b)

Figure 10: (Continued)



(c)

Figure 10: Confusion matrices for single alphabets classification. (a) G1, (b) G2, (c) G3

Table 11: Precision, recall, overall accuracy, and loss for G1 of single alphabets, shown in Fig. 10a

Group G1	Alif (SL1)	Ba (SL2)	Ta (SL3)	Tha (SL4)	Jeem (SL5)	Ha (SL6)	Dal (SL7)	Ra (SL8)	Sien (SL9)	Ya (SL10)	Wow (SL11)
Precision	80	77.8	50	62.5	100	57.1	25	55.6	50	83.3	100
Recall	88.9	77.8	55.6	55.6	100	44.4	44.4	55.6	33.3	55.6	100
Overall accuracy 72.66%											
Loss 1.6											

Table 12: Precision, recall, overall accuracy, and loss for G2 of single alphabets, shown in Fig. 10b

Group G2	Kha (SL1)	Dhal (SL2)	Za (SL3)	Sheen (SL4)	Dhad (SL5)	Ta (SL6)	Ghien (SL7)	Fa (SL8)	Kaf (SL9)	Ba (SL10)	Wow (SL11)
Precision	71.4	100	42.9	100	100	100	36	53.8	100	64.3	90
Recall	55.6	11.1	66.7	55.6	44.4	22.2	100	77.8	44.4	100	100
Overall accuracy 70.31%											
Loss 1.7											

Table 13: Precision, recall, overall accuracy, and loss for G3 of single alphabets, shown in Fig. 10c

Group G3	Sad (SL1)	Dhad (SL2)	Aieen (SL3)	Qaf (SL4)	Noon (SL5)	Hamzah (SL6)	Meem (SL7)	Wow (SL8)	Ha (SL9)	Lam (SL10)
Precision	77.8	87.5	50	69.2	100	100	85.7	75	66.7	100
Recall	77.8	77.8	66.7	100	66.7	100	66.7	100	66.7	55.6
Overall accuracy 77.48%										
Loss 1.3										

Table 14 summarizes of overall accuracy results according to all experiments and observations in the above three paragraphs. Though the classification model's prediction performance for Quranic words and disjoined letters was superior, ranging from 80.47% to 83.33%, it was somewhat lower for single Arabic letters, with a 73.5% average accuracy for all groups. Therefore, it may be important

to carry out further investigations to enhance the performance of their prediction, which can be a potential avenue for future research. Our findings are comparable to or even better than those of the mandarin language, as shown in Table 15, which may emphasize the effect of distinctions between the benchmarked languages in the lip-reading challenge. The transfer learning model we employed in our experiments was pre-trained utilizing the LRW dataset.

Table 14: All the experimental results of the classification model on the AQAND

Category	Quranic words	Disjoined letters	Single letters		
Accuracy	83.33%	80.47%	G1	G2	G3
			72.66%	70.31%	77.48%

Table 15: Comparison with the existing work

Dataset	The experiment content type	Language	Model		Accuracy
			Frontend	Backend	
The LRW	Word-level	English	ResNet-18	3 Layers GRU	83.7%
LRW1000		Mandarin			46.5%
AQAND (ours)		Arabic			83.3%

5 Conclusions and Future Work

In this paper, to the best of our knowledge, we provide the first dataset of the Arabic language based on the book *Al-Qaida Al-Noorania* (AQAND) comprising single letters, disjoined letters, and Quranic words as video data samples captured from three different viewpoints (0°, 30°, and 90°). We use the new, proposed AQAND to train a CNN-based deep learning model using transfer learning for visual speech recognition (known as lip-reading). For disjoined letters, our model achieved an accuracy of 80.5% while achieving an accuracy of 77.5% for single letters. However, the experimental results showed that the performance of lip-reading for Quranic words in recognizing completely unseen test data samples achieved the highest accuracy of 83.33%.

Three future directions that we see are as follows: first, an additional CNN deep learning model can be trained as an initial step before the classification model to split and sort the single alphabets of each viseme into subgroups instead of the manually implemented splitting, to be then fed into the classification model for prediction. Second, the experimental work on the dataset of AQAND can be expanded by examining the proposed deep learning model on the remaining two replications of video data captured in 30° and 90° viewpoints and comparing their performance results from different aspects and in multiple scenarios. Third, the serving and contributing to the field of continuous speech recognition at a sentence level, as we intend to add complete Quranic verses to the dataset in the future. As such, establishing the basic artificial intelligence capability of Arabic Quranic lip-reading can be considered as a preface for further developments in recognition of continuous recitation of Holy Quran verses, which contributes to effectively employing the field of artificial intelligence in teaching the correct recitation of the Holy Quran.

Acknowledgement: The authors would like to thank King Abdulaziz University Scientific Endowment for funding the research reported in this paper. They would also like to thank the Islamic University of Madinah for its extensive help in the data collection phase.

Funding Statement: This research was supported and funded by KAU Scientific Endowment, King Abdulaziz University, Jeddah, Saudi Arabia.

Availability of Data and Materials: The AQAND dataset proposed in this work is available at (https://forms.gle/x5tQcDLeZUwJyG789).

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

References

[1] N. Alhaqani, "Al-qaida Al-noorania," *Al-Furqan Center for Quran Learning*, vol. 1, no. 1, pp. 36, 2010.

[2] A. Sagheer, T. Naoyuki and R. Taniguchi, "Arabic lip-reading system: A combination of hypercolumn neural network model with hidden Markov model," *Proceedings of International Conference on Artificial Intelligence and Soft Computing*, vol. 2004, pp. 311–316, 2004.

[3] D. Pascal, "Visual speech recognition of modern classic Arabic language," in *2011 Int. Symp. on Humanities, Science and Engineering Research*, Kuala Lumpur, Malaysia, vol. 1, pp. 50–55, IEEE, 2011.

[4] J. S. Chung and A. Zisserman, "Learning to lip read words by watching videos," *Computer Vision and Image Understanding*, vol. 173, no. 5, pp. 76–85, 2018.

[5] F. Tao and C. Busso, "End-to-end audio-visual speech recognition system with multitask learning," *IEEE Transactions on Multimedia*, vol. 23, pp. 1–11, 2020.

[6] F. Xue, T. Yang, K. Liu, Z. Hong, M. Cao *et al.,* "LCSNet: End-to-end lipreading with channel-aware feature selection," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 18, no. 4, pp. 27, 2022.

[7] T. Stafylakis, M. H. Khan and G. Tzimiropoulos, "Pushing the boundaries of audio-visual word recognition using residual networks and LSTMs," *Computer Vision and Image Understanding*, vol. 176, pp. 22–32, 2018.

[8] H. Wang, G. Pu and T. Chen, "A lip reading method based on 3D convolutional vision transformer," *IEEE Access*, vol. 10, pp. 77205–77212, 2022.

[9] Y. Lu, H. Tian, J. Cheng, F. Zhu, B. Liu *et al.,* "Decoding lip language using triboelectric sensors with deep learning," *Nature communications*, vol. 13, no. 1, pp. 1–12, 2022.

[10] S. Jeon and M. S. Kim, "End-to-end sentence-level multi-view lipreading architecture with spatial attention module integrated multiple CNNs and cascaded local self-attention-CTC," *Sensors*, vol. 22, no. 9, pp. 3597, 2022.

[11] D. Tsourounis, D. Kastaniotis and S. Fotopoulos, "Lip reading by alternating between spatiotemporal and spatial convolutions," *Journal of Imaging*, vol. 7, no. 5, pp. 91, 2021.

[12] D. Feng, S. Yang, S. Shan and X. Chen, "Learn an effective lip reading model without pains," arXiv preprint arXiv: 2011.07557, 2020.

[13] J. S. Chung and A. Zisserman, "Lip reading in the wild," *Asian Conference on Computer Vision*, vol. 13, no. 2, pp. 87–103, 2016.

[14] S. Yang, Y. Zhang, D. Feng, M. Yang, C. Wang *et al.,* "LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild," in *2019 14th IEEE Int. Conf. on Automatic Face & Gesture Recognition (FG, 2019)*, Lille, France, vol. 1, pp. 1–8, 2019.

[15] E. Egorov, V. Kostyumov, M. Konyk and S. Kolesnikov, "LRWR: Large-scale benchmark for lip reading in russian language," arXiv preprint arXiv: 2109.06692, 2021.

[16] S. Jeon, A. Elsharkawy and M. Sang Kim, "Lipreading architecture based on multiple convolutional neural networks for sentence-level visual speech recognition," *Sensors*, vol. 22, no. 1, pp. 72, 2021.

[17] Ü. Atila and F. Sabaz, "Turkish lip-reading using Bi-LSTM and deep learning models," *Engineering Science and Technology, An International Journal*, vol. 1, no. 35, pp. 101206, 2022.

[18] Y. Lu, Q. Xiao and H. Jiang, "A chinese lip-reading system based on convolutional block attention module," *Mathematical Problems in Engineering*, vol. 2021, no. 12, pp. 1–12, 2021.

[19] N. Ziafat, A. Hafiz Farooq, F. Iram, Z. Muhammad, A. Alhumam and R. Kashif, "Correct pronunciation detection of the Arabic alphabet using deep learning," *Applied Sciences*, vol. 11, no. 6, pp. 2508, 2021.

[20] A. Asif, H. Mukhtar, F. Alqadheeb, A. Hafiz Farooq and A. Alhumam, "An approach for pronunciation classification of classical Arabic phonemes using deep learning," *Applied Sciences*, vol. 12, no. 1, pp. 238, 2021.

[21] P. Damien, N. Wakim and M. Egea, "Phoneme-viseme mapping for modern, classical Arabic language," in *2009 Int. Conf. on Advances in Computational Tools for Engineering Applications*, Zouk Mosbeh, Lebanon, vol. 1, pp. 547–552, 2009.

[22] F. Z. Chelali, K. Sadeddine and A. Djeradi, "Visual speech analysis application to Arabic phonemes," *Special Issue of International Journal of Computer Applications (0975-8887) on Software Engineering, Databases and Expert Systems-SEDEXS*, vol. 102, no. 1, pp. 29–34, 2012.

[23] T. Altalmas, M. Ammar, S. Ahmad, W. Sediono, J. Momoh *et al.,* "Lips tracking identification of a correct Quranic letters pronunciation for Tajweed teaching and learning," *IIUM Engineering Journal*, vol. 18, no. 1, pp. 177–191, 2017.

[24] L. Elrefaei, T. Alhassan and S. Omar, "An Arabic visual dataset for visual speech recognition," *Procedia Computer Science*, vol. 163, no. 10, pp. 400–409, 2019.

[25] W. Dweik, S. Altorman and S. Ashour, "Read my lips: Artificial intelligence word-level Arabic lip-reading system," *Egyptian Informatics Journal*, vol. 23, no. 2, pp. 1–12, 2022.

[26] N. Alsulami, A. Jamal and L. Elrefaei, "Deep learning-based approach for Arabic visual speech recognition," *CMC-Computers, Materials & Continua*, vol. 71, no. 1, pp. 85–108, 2022.

[27] M. Y. El Amrani, M. H. Rahman, M. R. Wahiddin and A. Shah, "Building CMU Sphinx language model for the Holy Quran using simplified Arabic phonemes," *Egyptian informatics journal*, vol. 17, no. 3, pp. 305–314, 2016.

[28] S. Abed, M. Alshayeji and S. Sultan, "Diacritics effect on Arabic speech recognition," *Arabian Journal for Science and Engineering*, vol. 44, no. 11, pp. 9043–9056, 2019.

[29] Al-Kaf, M. Sulong, A. Joret, N. Aminuddin and C. Mohammad, "QVR: Quranic verses recitation recognition system using pocketsphinx," *Journal of Quranic Sciences and Research*, vol. 2, no. 2, pp. 35–41, 2021.

[30] M. Rafi, B. Khan, A. W. Usmani, Q. Zulqarnain, A. Shuja *et al.,* "Quran companion-A helping tool for huffaz," *Journal of Information & Communication Technology*, vol. 13, no. 2, pp. 21–27, 2019.