technologies

MDPI

Article

Cross-Attention Fusion of Visual and Geometric Features for Large-Vocabulary Arabic Lipreading

Samar Daou [1,*], Achraf Ben-Hamadou [1,2], Ahmed Rekik [1,3] and Abdelaziz Kallel [1,2]

1 SMARTS Laboratory, Technopark of Sfax, Sakiet Ezzit, Sfax 3021, Tunisia; achraf.benhamadou@crns.rnrt.tn (A.B.-H.); ahmed.rekik@crns.rnrt.tn (A.R.); abdelaziz.kallel@crns.rnrt.tn (A.K.)
2 Digital Research Center of Sfax, Technopark of Sfax, Sakiet Ezzit, Sfax 3021, Tunisia
3 ISSAT Institut Supérieur des Sciences Appliquées et de Technologie, Gafsa University, Sidi Ahmed Zarrouk University Campus, Gafsa 2112, Tunisia
* Correspondence: samar.daou@enis.tn

Abstract: Lipreading involves recognizing spoken words by analyzing the movements of the lips and surrounding area using visual data. It is an emerging research topic with many potential applications, such as human–machine interaction and enhancing audio-based speech recognition. Recent deep learning approaches integrate visual features from the mouth region and lip contours. However, simple methods such as concatenation may not effectively optimize the feature vector. In this article, we propose extracting optimal visual features using 3D convolution blocks followed by a ResNet-18, while employing a graph neural network to extract geometric features from tracked lip landmarks. To fuse these complementary features, we introduce a cross-attention mechanism that combines visual and geometric information to obtain an optimal representation of lip movements for lipreading tasks. To validate our approach for Arabic, we introduce the first large-scale Lipreading in the Wild for Arabic (LRW-AR) dataset, consisting of 20,000 videos across 100 word classes, spoken by 36 speakers. Experimental results on both the LRW-AR and LRW datasets demonstrate the effectiveness of our approach, achieving accuracies of 85.85% and 89.41%, respectively.

Keywords: lipreading; deep learning; LRW-AR; graph neural networks; Transformer; Arabic language

check for updates

Academic Editor: Pedro Antonio Gutiérrez

Received: 15 November 2024
Revised: 23 December 2024
Accepted: 26 December 2024
Published: 9 January 2025

Citation: Daou, S.; Ben-Hamadou, A.; Rekik, A.; Kallel, A. Cross-Attention Fusion of Visual and Geometric Features for Large-Vocabulary Arabic Lipreading. *Technologies* **2025**, *13*, 26. https://doi.org/10.3390/technologies13010026

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Lipreading is the technique of recognizing spoken words by analyzing the movements of the lips and surrounding area of a speaking person in a video. Its applications span across different domains such as assisting deaf people [1], analysis of criminal conversations [2], speaker identification [3], and human–machine interaction systems such as dictating instructions [4–6]. This technique proves especially valuable in challenging environments with noise or multiple speakers, where hearing is challenging, yet comprehension is essential [7]. Additionally, when combined with audio, lipreading can offer supplementary information that enhances the speech recognition system's performances, especially in noisy conditions [8].

Learning to read lips is challenging due to various factors. One main challenge is the similarity in lip movements for certain distinct characters in a language, such as 'p' and 'b' in English. Furthermore, lipreading encounters additional challenges, such as variations in facial appearance, accents, speaking speed and manner, as well as variations in facial

pose [9]. These factors further contribute to the complexity of lipreading and require robust training and advanced architectures to account for individual differences and enhance accuracy in lipreading systems.

In recent years, there has been growing interest in the research field of lipreading, mainly driven by the advancements in deep learning techniques and the availability of parallel computation acceleration hardware. Deep learning-based lipreading systems have demonstrated superior performance compared to earlier approaches based on handcrafted image descriptors for many applications [10].

Deep learning-based lipreading systems encounter several challenges. It is a fact that the availability of learning datasets significantly influences the development of deep learning network architectures and language-specific lipreading systems, considering the substantial variations in pronunciation, phonemes, and lip movements between languages [11]. In this sense, several lipreading datasets have been created in different languages to fulfill this requirement. For instance, an English lipreading dataset [12] comprises 500,000 instances of spoken words, while a Mandarin dataset [13] contains 286 Chinese syllables represented by more than 1 million character occurrences. More recently, a Russian dataset [14] was introduced, consisting of over 117,500 samples of 235 words. In this work, we specifically focus on the Arabic language. Despite efforts made in this field, it is worth noting that Arabic, spoken by over 274 million people worldwide (https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide/, accessed on 10 September 2024), lacks an existing large-scale lipreading dataset or a dedicated system designed to this language.

Moreover, using only visual features computed on the cropped mouth region for lipreading has big challenges. These features include a lack of contextual information, inherent ambiguity and variability in lip patterns, limited discriminative power, and sensitivity to environmental factors. However, these limitations can be effectively mitigated by incorporating facial landmarks associated with video frames. By explicitly incorporating facial landmarks, lipreading systems gain access to additional precise geometric information regarding fine-grained lip movements and shape variations, which enables better differentiation of visually similar phonemes. While some existing methods use both video frames and landmarks for lipreading (e.g., [15,16]), they often rely on basic combination methods such as feature concatenation that may limit their ability to effectively capture essential mutual information. In this work, we focus specifically on the fusion aspect and introduce a sophisticated cross-attention mechanism for optimally combining visual and landmark features in our lipreading system.

To summarize, our main contributions are as follows:

- We introduce an efficient cross-attention fusion mechanism that merges image and facial landmarks for lipreading. Our fusion strategy aims to capture a larger visual context and improve the combination of mouth appearance and landmark features.
- We present the first large-scale Arabic dataset for word-level lipreading, containing 20,000 videos for 100 different word classes.

All resources and additional details related to this work are available on the project page: https://crns-smartvision.github.io/lrwar (accessed on 27 December 2024).

The remainder of this paper is organized as follows: First, we provide an overview of related works. Afterward, we outline the architecture of our lipreading system, followed by a presentation of the proposed dataset, including details on data collection, annotation, and preprocessing methods. We then report the obtained experimental results. Finally, we conclude this paper by summarizing our findings and highlighting areas for future work.

2. Related Work

In this section, we start with an overview of publicly available lipreading datasets in multiple languages, with a specific focus on the Arabic language. Then, we review the most relevant deep learning-based lipreading approaches closely related to our work.

2.1. Lipreading Datasets

2.1.1. Large-Scale Lipreading Datasets

Despite many efforts to improve classification performance, lipreading networks are known for requiring extensive training data. Consequently, early attempts at deep learning-based methods faced a data shortage, leading to the increased popularity of databases with more samples per class. Table 1 summarizes the most popular publicly available datasets.

Table 1. Most well known large-vocabulary lipreading datasets.

Dataset	Language	Speakers	Task	Classes	Utterances	Duration	Year
GRID [17]	English	34	Sentences	51	34,000	~ 27 h	2006
LRW [12]	English	1000+	Words	500	400,000	~111 h	2016
VoxCeleb [18]	English	2000+	Sentences	1000	-	-	2017
LRS [19]	English	1000+	Sentences	17,428	118,116	~ 33 h	2017
MV-LRS [20]	English	1000+	Sentences	14,960	74,564	~ 20 h	2017
LRS2-BBC [21]	English	1000+	Sentences	-	8 M	-	2017
LRS3-TED [22]	English	5000+	Sentences	-	1,2 M	~ 400 h	2018
VoxCeleb2 [23]	English	6000	Sentences	-	1 M	-	2018
LRW-1000 [13]	Mandarin	2000+	Words	1000	-	-	2018
VoxSRC [24]	English	1000+	Sentences	-	19,154	-	2019
LRWR [14]	Russian	135	Words	235	-	-	2021

The GRID dataset, introduced in 2006 [17], was a pioneering attempt to construct a large-scale English lipreading dataset in terms of the number of samples per class but had limited vocabulary.

To address this limitation, new databases have been created more recently to provide a more extensive vocabulary and a larger number of samples. The most relevant among them are the LRW [12], LRS [19], and MV-LRS [20] databases.

The LRW and LRS databases are derived from BBC program recordings between 2010 and 2016.

The LRW dataset [12] contains sentences from more than 1,000 speakers and a vocabulary of 500 words that appear at least 800 times each (400,000 in total). Therefore, the large number of speakers helps to generalize to unseen speakers. The LRS dataset [19] consists of 17,428 distinct words combined in 118,116 statements, as well as the corresponding facial track. Also, the MV-LRS database [20] has been recorded from BBC programs; in contrast to LRW and LRS which only include front views, MV-LRS contains images from multiple viewing angles, ranging from 0 to 90 degrees. In addition, the LRS2-BBC [21] dataset is like an extension to LRS that consists of more than 8 million utterances. The LRS3-TED [22] dataset includes more than 400 h of video, taken from 5594 TED and TEDx talks in English.

LRW-1000 [13] is a naturally distributed lipreading dataset that contains 1000 classes with approximately 1.14 million Chinese character instances and spans 286 Chinese syllables. This database is the largest and the only publicly available Mandarin lipreading dataset at the moment. The LRWR dataset [14] is the largest available dataset in the Russian language for visual speech recognition and contains 235 classes and 135 speakers. While, VoxCeleb [18], VoxCeleb2 [23], and VoxSRC [24] are publicly available datasets, they consist of short clips of videos taken from unconstrained YouTube videos of thousands of celebrities talking in the wild.

2.1.2. Lipreading Datasets for Arabic Language

Compared to other languages, the unavailability of large-scale datasets for Arabic has been a barrier to developing efficient lipreading systems for the Arabic language.

The AVAS dataset [25] was introduced in 2013 as the first Arabic dataset in the field. It includes recordings of 50 speakers uttering digits, words, and phrases. However, this dataset lacks information on the number of classes or utterances included, limiting its utility for researchers.

In 2022, the RML dataset [26] was recorded with 73 speakers and included 10 classes. However, the exact number of utterances in the dataset is not specified. More recently, the ArabicVisual dataset [27] was released, involving 24 speakers and 14 words, for a total of 2400 utterances. Despite representing a significant improvement, both the RML and ArabicVisual datasets, like AVAS, remain limited in terms of the number of words and speakers.

Table 2 summarizes the aforementioned works that attempt to construct lipreading datasets for Arabic.

Table 2. List of available lipreading datasets for Arabic.

Name	Year	Speakers	Task	Classes	Utterances	Availability
AVAS [25]	2013	50	Digits, words, phrases	48	-	✗
RML [26]	2022	73	Words	10	-	✗
ArabicVisual [27]	2022	22	Digits, phrases	14	2400	✓

2.2. Lipreading Methods

The number of works addressing the lipreading task has increased significantly over the last decade, thanks to the availability of large-scale datasets and developments in deep learning techniques [28]. In the following sections, we begin by highlighting key works dedicated to the Arabic language, followed by an overview of lipreading methods designed for other languages.

2.2.1. Lipreading Methods for Arabic Language

Recent works on lipreading for the Arabic language have mainly focused on word recognition. One such system is the system proposed by Dweik et al. [26], which is a lipreading system employing deep learning techniques such as convolutional neural networks (CNNs), time-delayed CNN with long short-term memory (TD-CNN-LSTM), and time-delayed CNN with bidirectional LSTM (TD-CNN-BiLSTM). When tested on the RML dataset, it achieved an overall prediction accuracy of 82.84% on test data using the CNN model with RGB dataset. However, its applicability to larger vocabularies or more complex real-world scenarios is limited by the relatively small dataset size and lexicon. Additionally, while the use of CNN and LSTM architectures has proven effective, it does not leverage more recent advancements in deep learning, such as attention mechanisms and graph neural networks, which could potentially enhance performance further.

Additionally, Alsulami et al. [27] proposed their Arabic lipreading approach based on VGG-19 as a backbone model to extract visual features. The proposed approach achieved an accuracy of 94% for digit recognition, 97% for sentence recognition, and 93% for digit and sentence recognition on the ArabicVisual dataset. Similar to the approach by Dweik et al. [26], their method achieves prominent results on the ArabicVisual dataset. However, the dataset's limited scope, centered primarily on digits and a restricted range of sentence types, significantly reduces its applicability to broader, large-vocabulary lipreading tasks. Additionally, while VGG-19 has historically demonstrated effectiveness in visual feature extraction, its static architecture lacks the ability to model temporal dynamics inher-

ent in sequential data. This limitation poses a challenge for tasks that require detailed and continuous analysis of lip movements, such as large-scale speech recognition or complex sentence decoding.

2.2.2. Lipreading Methods for Other Languages

Several studies have explored different approaches for visual speech recognition, mainly focusing on the English language.

Gutierrez et al. [29] presented some word prediction models based on the MIRACL-VC1 dataset [30]. The data were processed by detecting and extracting the subject's face area in each video frame before entering the entire video sequence into the model. Their proposed architecture is based on deep multi-layer CNN models and LSTM networks, taking inspiration from LipNet [31]. The extensive experimental results demonstrated the importance of dropout, data augmentation, hyperparameter tuning, including visible and invisible validation splits, and batch normalization for the regularization of these models.

Petridis et al. [32] utilized a 3D convolutional neural network combined with ResNet-34 as the frontend to extract spatial and temporal features. This was followed by a bidirectional gated recurrent unit (Bi-GRU) network as the backend to capture the temporal dependencies across frames. On the English LRW dataset, it achieved 82% accuracy.

The lipreading approach proposed by Stafylakis et al. [33] consists of a 3D convolutional neural network and a residual network designed to capture relevant visual data from the input sequence for each timestamp. These representations are then passed into a bidirectional LSTM. The network was trained and tested on the LRW dataset using various configurations, with the best configuration achieving an accuracy of 83.0%, outperforming previous works presented in [12,20].

Zhao et al. [34] proposed a method that uses a 3D convolutional network combined with ResNet-18 as the frontend. Similar to [32], the backend comprises a Bi-GRU network. By introducing a mutual information-based loss function to enhance the backend, their method achieved an accuracy of 84.41% on the LRW dataset.

Ma et al. [35] have proposed a lipreading model that is specifically designed for recognizing isolated words. Their model is composed of a 3D convolutional network, 18 layers of residual networks, and a temporal convolutional network (TCN). The proposed model has achieved prominent results on the LRW and LRW-1000 datasets. In addition, they proposed a more resource-efficient approach by incorporating 3D convolution combined with Shufflenetv2 for efficient feature extraction in the frontend. And for the backend, they utilized a depthwise separable temporal convolutional network (TCN), achieving an accuracy of 88.5% on the LRW dataset.

Recently, the same authors introduced a new depth temporal convolutional layer head in [36], which improves the performance of the model with a significant reduction in computational cost. With this updated architecture, they achieved state-of-the-art results, with an accuracy of 88.6% and 46.6% on the LRW and LRW-1000 datasets, respectively.

Koumparoulis et al. [37] further advanced the field by integrating the EfficientNetV2-L network with a TCN and Transformer, achieving 89.52% on LRW while significantly reducing computational complexity. This demonstrated the potential of Transformer-based models in lipreading tasks.

In their method, Miled et al. [38] introduced a 2D CNN-based frontend paired with Bi-GRU, obtaining 90.38% accuracy on LRW. The standout feature of their work was the integration of a segmentation mechanism that plays a crucial role in improving performance. Without segmentation, the performance dropped significantly, achieving 85%.

The existing deep learning techniques for lipreading primarily focus on exploiting appearance and optical flow features from videos. However, these methods do not fully

utilize the potential of lip motion characteristics. In addition to appearance and optical flow, the dynamic mouth contour conveys valuable complementary information. Unfortunately, the modeling of dynamic mouth contours has not received sufficient attention compared to appearance and optical flow.

To bridge this gap, Sheng et al. [15] proposed an Adaptive Semantic-Spatio-Temporal Graph Convolution Network (ASST-GCN). By leveraging the architecture of graph convolutional networks, ASST-GCN captures the inherent relationships between the points of the mouth contour and their temporal dynamics. To combine the complementary information from appearance and mouth contours, they introduce a two-stream visual front-end network that simultaneously processes both types of data to extract distinctive features. And lately, Daou et al. [16] proposed a two-stream deep learning architecture. This architecture aims to integrate visual and geometric features extracted from the lips and surrounding area. However, while the fusion of information from the two previous works [15,16] is straightforward and relies on basic concatenation, it may limit the architecture's efficiency in capturing mutual information between the merged streams.

2.2.3. Attention Mechanisms for Lipreading

Recently, attention mechanisms have gained significant attraction in different domains [39] due to their ability to focus on informative parts of the input data. In the context of lipreading, self-attention mechanism has been explored for modeling temporal relationships within an input video sequence. For instance, Lohrenz et al. [40] propose a self-attention based architecture for lipreading, enabling the model to focus on relevant features extracted across the video sequence. Differently, cross-attention mechanisms have shown promising results in different applications and domains, especially for leveraging complementary additional modalities.

Nagrani et al. [41] applied cross-attention mechanisms in their multi-modal speech recognition system. Their approach showed significant improvements in recognizing speech under challenging conditions by effectively combining audio and visual signals, highlighting the power of cross-attention in enhancing speech recognition.

In natural language processing, Tsai et al. [42] proposed a multi-modal Transformer network that employs cross-attention to fuse audio, visual, and textual information. Their model achieved state-of-the-art results on several multi-modal benchmarks, showcasing the effectiveness of cross-attention in handling diverse data sources.

Furthermore, in their work, Rajan et al. [43] investigated the use of cross-attention for multi-modal emotion recognition. They found that cross-attention mechanisms can outperform self-attention by better capturing the interactions between visual and auditory features. This results in an improvement of the accuracy of emotion recognition systems, underscoring the potential of cross-attention in effectively merging different types of data.

2.3. Our Method's Positioning

To define lip movement, we propose a model that takes advantage of cross-modal properties. It employs a cross-attention mechanism to effectively align visual data with facial landmarks. This alignment enhances the fusion of diverse features, resulting in improving the model's performance. This approach contrasts with single attention fusion, which may not fully capture the complex dependencies and interactions between modalities. Although previous works using single attention mechanisms, such as [39], have shown potential, our results reveal that the cross-attention offers a significant performance enhancement.

To deliver high performance, we propose employing convolutional neural networks, including 3D and 2D convolutional models for extracting visual features and graph neural

networks to extract geometric features, along with temporal convolutional networks that have been highly effective in recent deep learning-based lipreading systems.

Motivated by this success in many languages, we propose the integration of these techniques into Arabic lipreading systems and study its potential. This adaptation could considerably improve the performance of computer-assisted communication applications in this language.

3. Proposed Approach

As depicted in Figure 1, our approach consists of five key steps. The first step involves video preprocessing, wherein we crop the mouth region of T frames from the input video sequence and acquire corresponding facial landmarks. This step yields both visual and geometric data streams. Afterward, a visual-feature network is employed to extract a sequence of T feature vectors for the visual data stream, and a geometric-feature network to encode the variation in lips contour delivered by facial landmarks. Moving forward, a FusionNet network is used to combine these two sequences into a single sequence of T feature vectors. The final step outputs the predicted speech by applying a sequence decoder network. Detailed descriptions of each module involved in these steps are provided below.

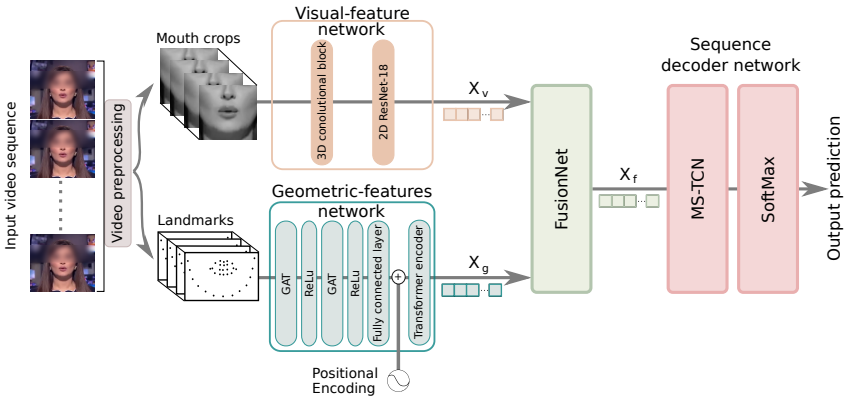


Figure 1. Outline of the proposed approach. Video preprocessing: crops the mouth region from the input video sequence and obtain the corresponding facial landmarks. Visual-feature network: extracts relevant information from the preprocessed data. Geometric-feature network: encodes lip contour variation delivered by facial landmarks. FusionNet network: fuses the encoded features. Sequence backend network: based on a multi-scale temporal convolutional network (MS-TCN) to encode temporal variation and classify the input video sequence.

3.1. Input Data Preprocessing

Assuming that a face detector is first applied to the input video to broadly localize the talking face in all input frames, we applied a generic facial landmark detector [44] to identify 68 standard landmarks on every input frame. Other landmark detection methods could also be considered for this study, as long as they cover landmarks on the lips. We use these landmarks to align and resize the frames, resulting in a mouth region of 96×96 pixels to ensure its position is nearly centered in the extracted images. Additionally, these cropped patches are converted to grayscale and normalized by subtracting the mean and dividing by the standard deviation values. As a data augmentation process during the training phase, an additional cropping is applied to obtain 88×88-size patches at random positions. For the validation and testing phases, a central cropping of 88×88-size patches is applied instead. On the other hand, the extracted landmarks undergo the same affine transformations used for frame preprocessing to ensure they are represented in the same coordinate system as the extracted image patches, thereby geometrically aligning the patches and landmarks. This

preprocessing step yields two streams of data: the extracted images of the mouth region, and the sequence of extracted facial landmarks.

3.2. Visual-Feature Network

The visual-feature network encodes the appearance and movement of the lips over the extracted image sequence. It consists of a single 3D convolutional block followed by a modified 2D ResNet-18 architecture [45]. The 3D convolutional layer takes as input a sequence of T consecutive grayscale images of size 96×96. It is a sequence of 64 kernels of size $5 \times 7 \times 7$, followed by batch normalization, ReLU activation, and max pooling with a kernel size of $1 \times 3 \times 3$. This initial network outputs spatio-temporal features with dimensions of $T \times 64 \times 24 \times 24$. These spatio-temporal features are then fed into the ResNet-18 architecture, which consists of four stages of convolutional blocks. Each stage contains several residual blocks, with the number of feature channels progressively increasing from 64 in the first stage, to 128 in the second, 256 in the third, and finally 512 in the fourth stage. This hierarchical feature extraction process enables the network to capture increasingly complex patterns and representations of the lip movements. After passing through the ResNet-18 blocks, the features undergo adaptive average pooling, which reduces each feature map to a single value, resulting in a feature vector. The final output of the visual-feature network are the visual features $X_v \in \mathbb{R}^{T \times d}$, where $d = 512$. These features serve as a compact and informative representation of the visual input, capturing both spatial and temporal dynamics crucial for the subsequent stages of lipreading.

3.3. Geometric-Feature Network

Similar to the visual-feature network, the geometric-feature network encodes the appearance and movement of the lips. However, the geometric features are extracted based on an explicit representation of the mouth region using the sequence of extracted landmarks. Let us define N as the number of used landmarks per frame in this network. The input size of the geometric-feature network is $T \times N \times 2$, as each landmark is represented by its 2D coordinates.

To encode the spatial arrangement between the facial landmarks, each frame represents a graph, and each landmark is a node within the graph. We apply the K-nearest neighbors algorithm to create the edges between the nodes, with k set to 5. The network includes two graph attention network (GAT) [46] layers: the first layer transforms input graphs from 2 to 16 channels per node; and the second layer further expands them to 64 channels, followed by a fully connected layer to output a feature vector of 512 per graph. Moreover, to encode the temporal variation in landmarks over the frames, we consider a positional encoding of size 512 to be added to the T vectors and fed into the sequence $T \times 512$, which is then processed by a Transformer encoder block of a single layer with 8 heads and 1024 as the dimensions of the feed-forward network model, leading to the geometric feature tensor $X_g \in \mathbb{R}^{T \times d}$ of the same size as the visual features.

3.4. FusionNet Network

The proposed FusionNet is based on a multi-head attention mechanism to aggregate the fused features from both the visual and geometric features extracted from their corresponding networks. We keep the original notation as in the original description of the Transformer architecture [47]. As illustrated in Figure 2, it consists of two multi-head attention modules that exchange their K (key) features. The idea behind this design is to integrate features from both modalities. After applying the multi-head attention unit involving h heads, we concatenate the obtained features into one single feature map, de-

noted by $X_f \in \mathbb{R}^{T \times d_f}$ with $d_f = 1024$. The multi-modal fusion process is mathematically represented in Equation (1):

$$X_f = \text{Concatenation}\big(\text{Attend}_\text{V}(X_\text{V}, X_\text{G}), \text{Attend}_\text{G}(X_\text{G}, X_\text{V})\big) \tag{1}$$

where X_f represents the enriched feature representation resulting from the fusion process, incorporating two distinct multi-head attention mechanisms dedicated to cross-modal interactions:

- $\text{Attend}_\text{V}(X_\text{V}, X_\text{G})$ represents the attention mechanism applied to the "Vision" input (X_v) attending to the "Landmark" input (X_G), capturing the interplay of information between modalities.
- $\text{Attend}_\text{G}(X_\text{V}, X_\text{G})$ signifies the attention mechanism applied to the "Landmark" input (X_G) attending to the "Vision" input (X_V), facilitating knowledge exchange across modalities.

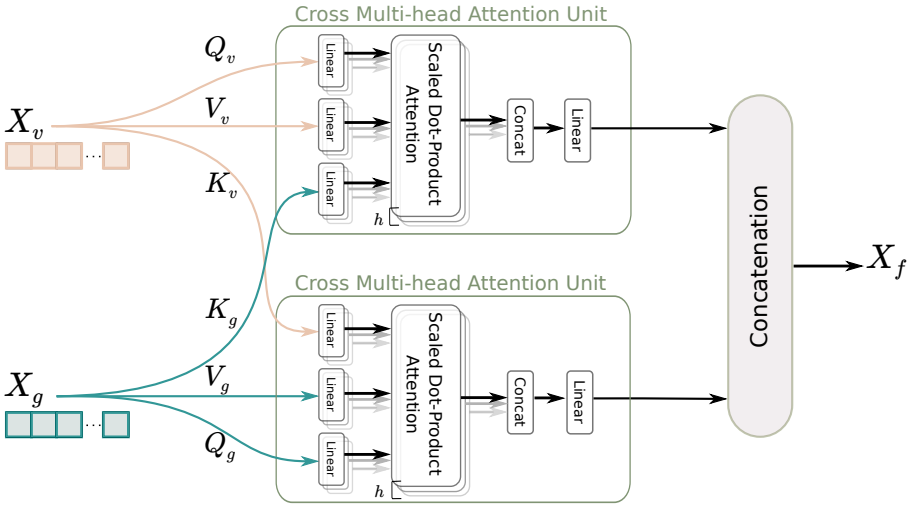


Figure 2. Architecture of FusionNet, consisting of two cross multi-head attention units with the same number of heads (h). The first unit processes visual features X_v, while the second handles geometric features X_g. Each unit takes as input a query (Q) and a value (V) derived from its respective modality features, while exchanging the input key (K) with the alternate modality.

3.5. Sequence Decoder Network

This sequence decoder network is based on the temporal convolution network (TCN) since it has been shown to significantly improve performance on word-level lipreading tasks [35]. Figure 3 presents a detailed representation of the proposed architecture. It contains multi-scale dilated TCN (MS-TCN) layers, where each layer comprises multiple TCN branches, which are 1D convolutions designed for the temporal domain, with different kernel sizes capturing information across various temporal scales. Our sequence decoder network consists of four layers of MS-TCN with dilation values of 1, 2, 4, and 8, respectively. Each layer is composed of three TCN units using kernel sizes of 3, 5, and 7. The outputs from these branches are concatenated and followed by a fully connected layer and a SoftMax layer, allowing the network to integrate information across multiple temporal scales for an improved spoken word prediction.

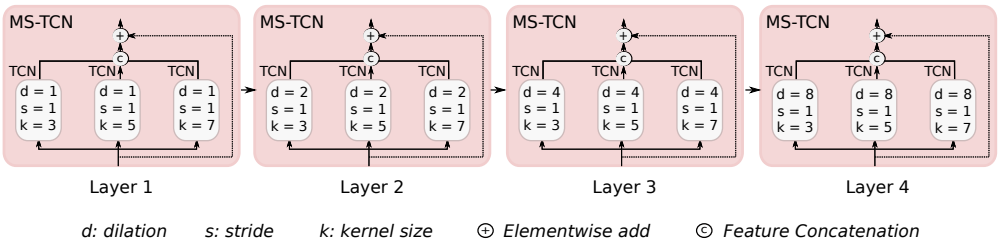


Figure 3. Sequence decoder network with four layers of multi-scale expanded TCNs with 1, 2, 4, and 8 dilation values, respectively. Each layer is composed of three TCN units with 3, 5, and 7 kernel values.

4. LRW-AR Dataset

In this section, we present an overview of the procedures followed to collect and prepare our large-scale dataset for Arabic lipreading. The fundamental concept was to gather videos of individuals speaking in Arabic in front of cameras from the YouTube platform, which is a rich source of diverse and natural language data. The process involved several stages, including video scrapping, shot boundary detection, face detection and tracking, data cleaning, annotation, video split, and data cropping, as illustrated in Figure 4. A diverse set of Arabic videos was collected and preprocessed to extract the frames of interest for lipreading. The frames were then annotated with the corresponding spoken words to create a ground truth for training and evaluation purposes. Finally, the annotations were visually validated to ensure their accuracy and consistency. Each of these steps will be described in detail in the following sections. The LRW-AR dataset is released under a BY-NC-ND license, and is accessible through the following link: https://osf.io/rz49x (accessed on 27 December 2024).

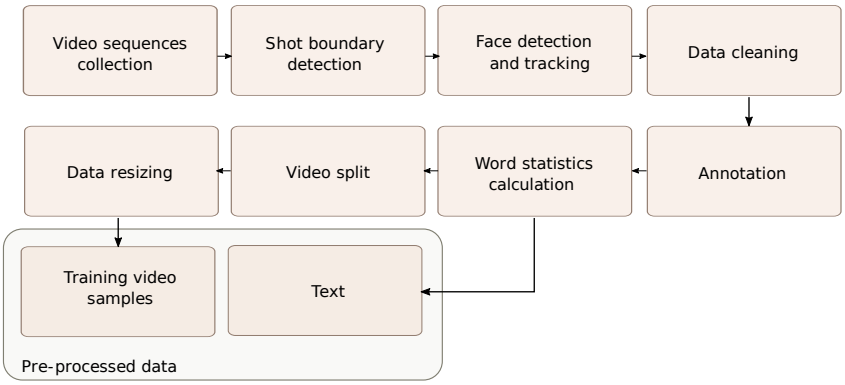


Figure 4. Pipeline to generate the LRW-AR dataset.

4.1. Data Collection and Preprocessing

4.1.1. Video Scraping

All videos were scraped from YouTube by selecting TV channel playlists that contained Arabic-language TV shows, news programs, and interviews. The playlists were chosen to cover a variety of speakers, genders, and accents.

4.1.2. Shot Boundary Detection

In this stage, we used a HOG-based scene detection algorithm [48] to detect scene changes in the collected videos, and they were split into shorter segments based on scene transitions. This allowed us to focus on the segments where the speaker was present and speaking.

4.1.3. Face Detection and Tracking

To ensure that our dataset only includes video sequences with a single speaker, the HOG-based face detection method [48] was applied to each raw video frame. Then, all face detections belonging to the same person were grouped across frames using a KLT tracker [49]. This filtering process helps to eliminate any potential confusion that may arise from having multiple speakers in a single video sequence. After filtering, we sampled all collected video segments into fixed-length sequences of 1.2 s each.

4.1.4. Data Cleaning

To ensure the dataset's quality, a manual inspection and cleaning process was carried out, ensuring that only videos featuring a single speaker speaking directly in front of the camera were included.

4.1.5. Annotation

For dataset annotation, we utilized the Vosk speech recognition system (https://alphacephei.com/vosk/, accessed on 10 September 2024) to automatically transcribe the spoken words in the video sequences into Arabic text. The use of Vosk not only facilitated transcriptions but also established boundaries, indicating the start and end of each spoken word within the text. While we anticipated a perfect alignment between text and audio data, there were some instances where the transcription was incorrect. To rectify this, manual correction of these errors was performed during the final phase of visual validation and dataset quality control.

4.1.6. Video Split

To generate a varied collection of Arabic words, statistical analysis was conducted on the transcriptions to identify the most frequently occurring words. This approach allowed us to focus on the prevalent words in Arabic speech. Subsequently, each video segment was divided into shorter clips, with each clip representing a single word. This procedure facilitated the development of an extensive dataset encompassing a wide array of Arabic words.

4.1.7. Data Cropping

In this step, we standardized all videos to a consistent size of 256 × 256 pixels. This uniformity enhances overall consistency and simplifies comparisons across different video segments. We cropped each video to focus solely on the speaker's face, minimizing any extraneous information and highlighting the lip movements of the speaker (see Figure 5). This enables us to achieve a more precise and comprehensive understanding of visual characteristics.

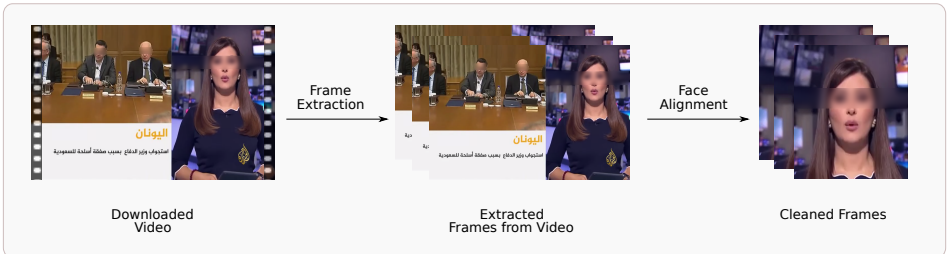


Figure 5. Frame extraction and face alignment process.

The resulting dataset contains a diverse set of Arabic words spoken by various speakers in different contexts, which can be used for lipreading research and the development of Arabic speech recognition systems. By sharing the methodology, the purpose is to

facilitate the creation of similar datasets for other languages and cultures and promote the development of more inclusive visual speech recognition systems.

4.2. Dataset Statistics

The proposed dataset consists of audio-visual data collected from YouTube videos from 2008 until 2022. The dataset contains 100 classes, each associated with a distinct word, and involves a total of 36 speakers (see Figure 6). Notably, the speakers are talking spontaneously, presenting additional challenges for lipreading systems, including differences in pronunciation, intonation, variations in face pose, and background noise. Each word is uttered 200 times, resulting in a total of 20,000 video samples in the dataset.



Figure 6. A sample of speakers in LRW-AR dataset.

Furthermore, each video sequence in the dataset corresponds to a 1.2 s audio clip captured at a frame rate of 25 frames per second with a video resolution of 256 × 256. In these sequences, the ground truth word is spoken in the middle, often surrounded by other words or sub-words. This characteristic adds complexity to the lipreading task by simulating more natural speaking conditions. Each video also comes with an annotation file that specifies the start and end times of the spoken words, ensuring precise alignment for training and evaluation purposes. The dataset was partitioned into training, validation, and testing sets. Specifically, 80% of the samples were designated for training, 10% for testing, and another 10% for validation. To provide a clearer understanding, Figure 7 showcases samples from the LRW-AR dataset. This figure displays three videos from three different classes, spoken by different speakers, with clear indications of where each corresponding word starts and ends.

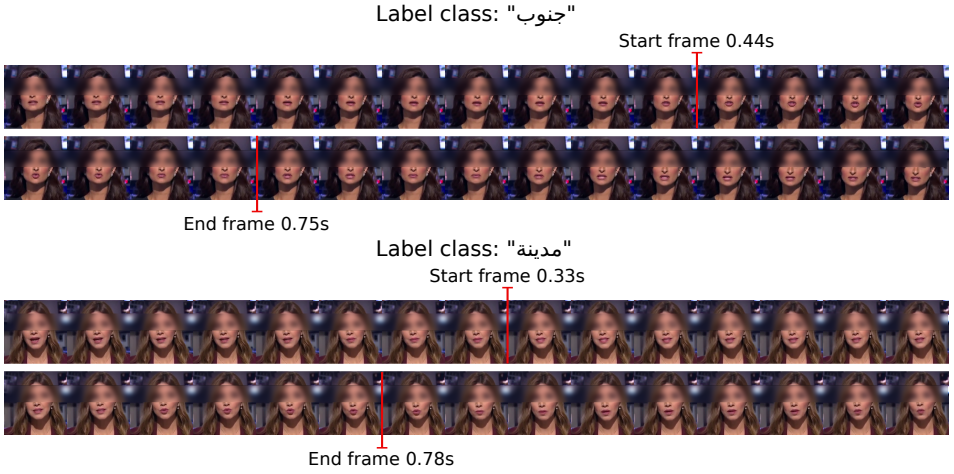


Figure 7. *Cont.*

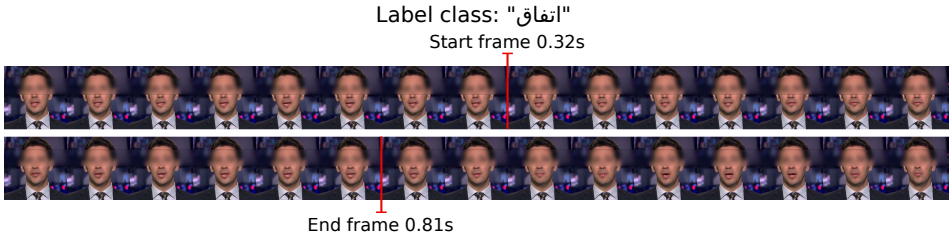


Figure 7. Three different samples from the LRW-AR dataset. The starting and ending frames for each uttered word are precisely indicated in every sample in the dataset.

5. Experiments and Results

In this section, we present various experiments conducted to assess the performance of the proposed lipreading system and its parameters. First, we provide details about our implementation and the datasets used in our experiments. Then, we conduct various experiments as part of our evaluation to demonstrate the precision and efficacy of our system, with a particular focus on lipreading for the Arabic language.

To ensure a more comprehensive comparison, we extended our evaluation to include the English LRW dataset. This allowed us to benchmark our method against other state-of-the-art lipreading techniques. This also enabled further validation of the LRW-AR dataset. While our focus remained on the Arabic language system, we conducted experiments examining the impact of different feature configurations, varying numbers of facial landmarks, and cross-attention heads on system accuracy. Finally, we present a comprehensive analysis outlining the strengths and weaknesses of our lipreading system, along with identified areas for potential improvement.

5.1. Implementation Details

Our implementation is based on the PyTorch library [50]. In all forthcoming experiments discussed in the remaining section, we trained all architecture configurations using an NVIDIA Titan V GPU with 12 GB of RAM. The training process was carried out with a mini-batch size of 32 for a total of 2000 epochs. The AdamW optimizer [51] was employed with an initial learning rate of 3×10^{-4}. We used a cosine annealing approach to reduce the learning rate without requiring a warm-up phase. Additionally, for all trials, we incorporated variable-length augmentation [36]. The code source is available through the following link: https://github.com/crns-smartvision/lrwar (accessed on 27 December 2024).

5.2. Datasets

In addition to our Arabic LRW-AR dataset, we conducted experiments using the LRW [12] and the ArabicVisual [27] datasets, commonly used benchmark datasets for lipreading. It is worth mentioning that ArabicVisual is currently the only lipreading dataset available for the Arabic language, whereas other datasets, such as the AVAS and RML datasets are not publicly available. To assess the performance of our lipreading model on the ArabicVisual dataset, we employed a random split of 75% for training, 15% for validation, and 10% for testing, as outlined in [27]. Accuracy was utilized as the evaluation metric, representing the percentage of correctly classified videos.

5.3. Arabic Lipreading Performance Evaluation

To evaluate the effectiveness of the proposed method, we conducted experiments using the LRW-AR dataset, comparing our method with state-of-the-art techniques. These experiments aim to demonstrate our model's performance in word-level lipreading tasks relative to these methods. We have selected the most relevant methods:

- Petridis et al. [32] utilized a 3D convolutional neural network (3D Conv) combined with ResNet-34 as the frontend to extract spatial and temporal features. This was followed by a bidirectional gated recurrent unit (Bi-GRU) network as the backend to capture the temporal dependencies across frames.
- Zhao et al. [34] proposed a method that employs a 3D Conv network along with ResNet-18 as the frontend. Similar to [32], the backend consisted of a Bi-GRU network.
- Ma et al. [35] introduced a method that uses 3D convolution combined with Shufflenetv2 for efficient feature extraction in the frontend. For the backend, a Depthwise Separable Temporal Convolutional Network (TCN) was used.
- Koumparoulis et al. [37] further advanced the field by integrating EfficientNetV2-L with a Transformer encoder (TCN + Tx).
- Miled et al. [38] introduced a 2D CNN-based frontend paired with Bi-GRU with the addition of a segmentation mechanism.

Table 3 summarizes the obtained results for the evaluated methods. The final row presents the results of our method, which integrates 3D Conv + ResNet-18 with GAT and Transformer (Tx) in the frontend, and employs MS-TCN + FusionNet in the backend. As shown, our method outperforms other state-of-the-art lipreading techniques, achieving the highest accuracy of 85.85% on the LRW-AR dataset compared to the accuracies of 78.60%, 79.85% , 83.23%, 83.94%, and 84.05% reported in [32,34,35,37,38] respectively.

Table 3. Accuracy performance of our method compared to state-of-the-art techniques on the LRW and LRW-AR datasets.

Method	Frontend	Backend	LRW	LRW-AR
[32]	3D Conv+ResNet-34	Bi-GRU	82%	78.60%
[34]	3D Conv+ResNet-18	Bi-GRU	84.41%	79.85%
[35]	3D Conv+Shufflenetv2	DepthwiseSeparableTCN	88.5%	83.23%
[37]	3D Conv+EfficientNetV2-L	TCN + Tx	89.52%	83.94%
[38]	2D CNN	Bi-GRU	90.38 %	84.05%
Ours	3D Conv+ResNet-18+GAT+Tx	MS-TCN+FusionNet	89.41%	85.85%

Additionally, the table reports the performance of our approach on the LRW dataset, where our method demonstrates competitive results, with an accuracy of 89.41%. Although this is slightly below the 89.52% achieved in [37] and the 90.38% in [38], it still outperforms other prominent works like [32] at 82%, [34] at 84.41%, and [35] at 88.5%.

Moreover, we compared our lipreading system with Asulami et al.'s work [27] on the ArabicVisual dataset based on the accuracy metric (percentage of correctly classified videos). Table 4 shows the obtained results for phrase classes only, digit classes only, and both digit and phrase classes.

Table 4. Comparative results of our lipreading system and the system developed by Alsulami et al. [27] on ArabicVisual dataset.

Method	Digits	Phrases	Digits and Phrases
Alsulami et al. [27]	94%	97%	93%
Ours	93.75%	100%	97.77%

Based on the obtained results, our approach appears to have achieved better accuracy than the Asulami et al. approach [27]. As outlined in Table 4, our approach achieved 100% accuracy for the 'phrases' and 97.77% for the 'digits and phrases' classes. However, Alsulami et al. reported accuracy rates of 97% and 93%, respectively, for the same classes. Nevertheless, our approach demonstrates competitive performance, with the Alsulami et

al. approach only marginally surpassing ours by 0.25% in digit recognition. These results highlight the limitations of the ArabicVisual dataset [27] due to its small scale in both size and vocabulary, emphasizing the need for a larger dataset that includes more challenging and representative data.

5.4. Ablation Study

5.4.1. Facial Landmark Selection

In this section, we conducted experiments to assess how the landmark subset selection affects the performance of our system. We evaluated the accuracy of our approach using the standard set of 68 face landmarks, encompassing all facial landmarks, and two subsets of landmarks: 33 points (covering the lower part of the face) and 20 points (covering the lips only). Figure 8 illustrates these three configurations. The obtained results for these three subsets of landmarks are presented in Table 5. As indicated in Table 5, employing only 20 landmark points yields the lowest accuracy, with 83.8%. Conversely, utilizing 68 landmark points resulted in improvement, with an accuracy of 84%. Furthermore, by employing 33 landmarks, we attained the highest accuracy of 85.85%.

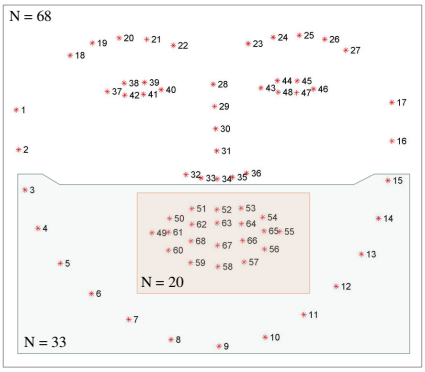
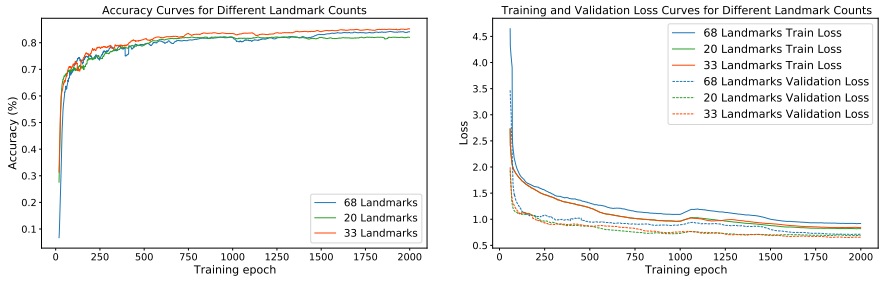


Figure 8. The process involves choosing *N* facial landmark points, which are points used to identify and analyze facial features. The total number of facial landmarks is 68, of which 33 are situated in the lower half of the face (marked by a blue box), and 20 are concentrated in the mouth region (marked by an orange box).

Table 5. Our proposed architecture's performance for the three subsets of selected landmarks.

	Visual Landmarks		
	20 Points	**33 Points**	**68 Points**
Accuracy	83.8%	85.85 %	84%

Figure 9 illustrates the curves for validation and training accuracy (a) and losses (b) across the varying numbers of employed landmarks. Notably, the optimal accuracy is attained when utilizing 33 landmarks. Moreover, the loss curves associated with 33 landmarks exhibit superior convergence speed and overall performance. These results highlight the effectiveness of FusionNet in leveraging landmarks to improve lipreading accuracy, where the 33-landmark configuration represents the optimal choice for our lipreading systems.



(a) (b)

Figure 9. Accuracy and loss curves: (**a**) validation and training accuracy curves, (**b**) validation and training loss curves.

5.4.2. Selection of Number of Heads in FusionNet

To enhance the performance of our lipreading system, we evaluated the number of heads (h) in the cross-attention fusion mechanism. Our goal was to investigate the influence of this parameter on our system's ability to recognize spoken words. We conducted evaluations with various values of h: 1, 2, 4, 8, and 16. Table 6 summarizes the performance metrics for different values of h.

Table 6. Performance metrics for various h values in cross-attention fusion.

	Number of Heads				
	$\boldsymbol{h = 1}$	$\boldsymbol{h = 2}$	$\boldsymbol{h = 4}$	$\boldsymbol{h = 8}$	$\boldsymbol{h = 16}$
Accuracy	83.2%	83.85%	84.60%	85.85%	82.25%

Starting with a single head $h = 1$, our system achieved an accuracy of 83.2%, indicating an acceptable baseline performance. As we raised the number of heads to $h = 2$, the accuracy increased a bit to 83.85%. This shows that a dual-headed approach slightly enhanced our system's performance. However, the most improvement in performance was achieved when $h = 4$, resulting in an accuracy of 84.60%. This indicates that a moderate level of parallelism and attention capacity is beneficial in decoding spoken words. On the other hand, the performance of our system improved, reaching an accuracy of 85,85% with $h = 8$. This configuration outperformed all other settings in capturing the complex dynamics of cross-modal interactions. However, increasing the number of heads to $h = 16$ led to a slight drop in accuracy to 82.25%, demonstrating that an excessive number of heads may add needless complexity. In summary, our thorough investigation demonstrates the critical role of the number of heads in the cross-attention fusion mechanism. The findings show that, among the evaluated configurations, $h = 8$ heads represents the best option, offering the highest level of precision in decoding both visual and geometric features.

5.4.3. Modality Fusion Comparison

In this experiment, our objective was to evaluate the performance of our lipreading system in the Arabic language using five distinct configurations: visual only (VO), landmarks only (LO), concatenated visual and landmark features (VL[concat]), fused features using single-modality attention (VL[SingleAtt]), and our full system configuration using FusionNet (VL[FusionNet]).

For the LO configuration, extracted landmarks were forwarded to the geometric-feature network. In the VO configuration, preprocessed data were input into the visual-feature network. In the combined configuration, we utilized three distinct architectures: the concatenated architecture VL[concat]; the VL[SingleAtt] architecture, where we used a sepa-

rate attention mechanism for each modality; and the FusionNet architecture VL[FusionNet], leveraging the synergy between visual and geometric features.

Table 7 presents a comparison of various network architectures and their performance on the LRW-AR dataset, demonstrating the effectiveness of different combinations of visual data and landmark configurations. The LO configuration with 68 landmarks achieved an accuracy of 65.45%, showing the value of detailed geometric features. The VO configuration performed significantly better with an accuracy of 83.45%, highlighting the importance of visual features. When combining visual data with landmarks, the VL[concat] approach saw the highest accuracy of 84.25% with 68 landmarks. The VL[SingleAtt] method, integrating visual data with 33 landmarks using a single attention mechanism, achieved a notable accuracy of 85.4%. However, the VL[FusionNet] method outperformed all other configurations, reaching the highest accuracy of 85.85% with 33 landmarks, showcasing the superior capability of FusionNet in effectively fusing visual and geometric features for enhanced lipreading accuracy.

To qualitatively evaluate our system's performance, we present the confusion matrices obtained by different configurations, underlined in Table 7. In Figure 10, the confusion matrices present five different configurations: landmarks only (LO), visual only (VO), concatenation architecture VL[concat], VL[SingleAtt], and VL[FusionNet] architecture, respectively. We note that the classes in the confusion matrices are sorted by the number of characters in the word, where all the words are arranged in ascending order based on their length (number of characters), from the shortest to the longest.

Table 7. Performance comparison of different network configurations and modalities on the LRW-AR dataset. The table presents the accuracy achieved with various combinations of visual data and the three subsets of landmarks (i.e., 20, 33, and 68). Underlined values indicate the best performance within a same network architecture across various modality combination settings, while bold values represent the best overall performance.

Network Architecture	Modality				Accuracy
	Visual	20 Landmarks	33 Landmarks	68 Landmarks	
LO	N/A	✓			59.75%
	N/A		✓		62.33%
	N/A			✓	<u>65.45%</u>
VO	✓	N/A	N/A	N/A	<u>83.45%</u>
VL[concat]	✓	✓			82.05%
	✓		✓		83.75%
	✓			✓	<u>84.25%</u>
VL[SingleAtt]	✓	✓			84.20%
	✓		✓		<u>85.4%</u>
	✓			✓	83.75%
VL[FusionNet]	✓	✓			83.8%
	✓		✓		**<u>85.85%</u>**
	✓			✓	84.0%

The obtained confusion matrices show that the VL[FusionNet] architecture consistently outperforms the other architectures, achieving the highest accuracy across all classes with different word lengths. Additionally, the VL[concat] and VL[SingleAtt] architectures demonstrate their robustness, especially for shorter word classes, but to a lesser extent.

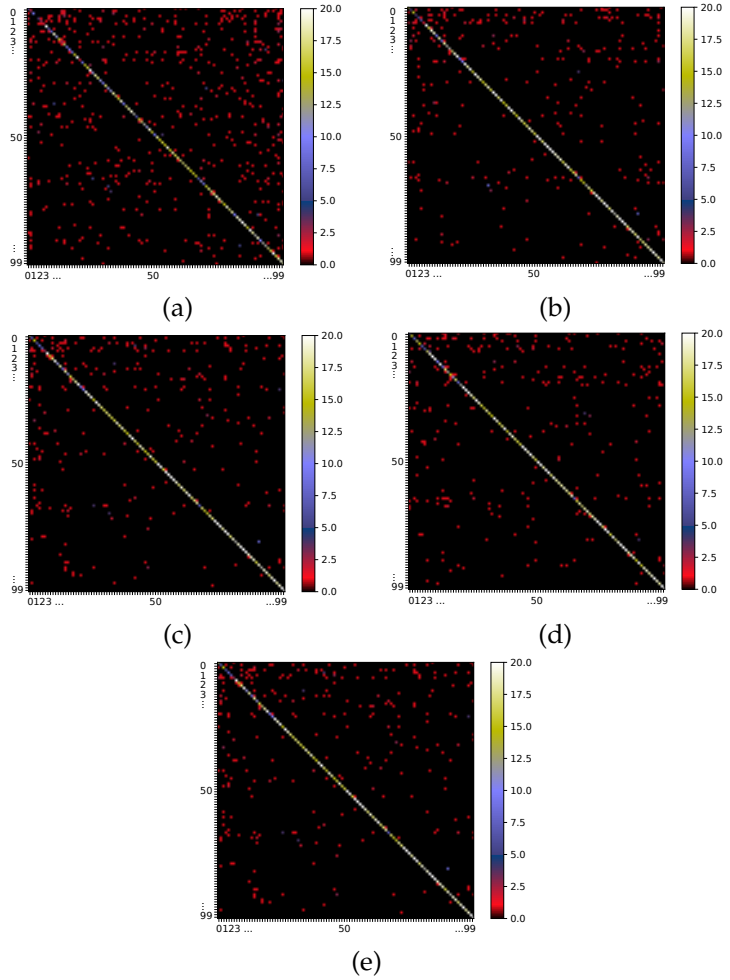


Figure 10. Confusion matrices for different models. (**a**) LO; (**b**) VO; (**c**) VL[concat]; (**d**) VL[SingleAtt]; (**e**) VL[FusionNet].

In contrast, the LO architecture demonstrates reasonable performance, achieving high accuracy for certain longer classes but struggling with shorter ones. Nevertheless, the VO architecture, despite surpassing the LO configuration, encounters difficulties with certain shorter classes but performs more effectively with longer ones.

In summary, the VL[FusionNet] architecture excels at recognizing both shorter and longer classes, demonstrating its adaptability and efficiency in handling a large vocabulary. These results highlight the importance of incorporating both visual and landmark information into the system, with FusionNet emerging as the most robust and accurate configuration for Arabic lipreading.

5.5. Discussion

Several key observations can be drawn from the experimental results. First, we conducted a comprehensive set of experiments to benchmark our approach against state-of-the-art methods in Arabic lipreading systems. Our proposed method demonstrated competitive performance, achieving an accuracy of 85.85%, which surpasses the results of existing methods, that reached 78.60%, 79.85%, and 83.23%, respectively. These results highlight the effectiveness of our approach in improving the accuracy of Arabic lipreading tasks.

For the ablation study, we conducted several experiments to demonstrate the complementarity between geometric features and visual cues, as well as the effectiveness of the proposed cross-attention mechanism in leveraging both modalities for optimal per-

formance. Specifically, we evaluated the impact of different subsets of landmarks (i.e., 20, 33, and 68 landmarks) and the number of heads in the multi-head cross-attention mechanism. Our experiments revealed that employing eight heads resulted in the highest accuracy, reaching 85.85%, as previously discussed. These results showed the need for a cross-attention mechanism, rather than straightforward feature concatenation and standard attention mechanisms, to achieve the best fusion of geometric and visual features.

In addition to validating our lipreading method, the experiments also aimed to evaluate the constructed LRW-AR dataset. We selected state-of-the-art methods based on the availability of their official implementations, even though these methods were not originally designed for Arabic lipreading. We benchmarked these methods on LRW-AR, and the consistent performance across various models highlights the dataset's robustness and demonstrates that it is not biased toward specific architectures. Moreover, when comparing the performance of the state-of-the-art methods on LRW-AR to those obtained on the English LRW dataset, we observed comparable results, which further validates the quality and design of our LRW-AR dataset. Furthermore, the uniqueness of LRW-AR lies in being the first large-scale lipreading dataset specifically designed for the Arabic language, marking a significant contribution to the field.

We also intentionally tested another available Arabic lipreading dataset, ArabicVisual [27], to highlight the size limitations. Our method achieved 100% accuracy in phrase recognition on this dataset, indicating that it no longer presents a significant challenge. This underscores the need for larger datasets with more extensive vocabularies. While lipreading datasets for English have received considerable attention, LRW-AR is the first large-scale lipreading dataset specifically for Arabic. It represents a significant advancement in the field and could be continuously expanded in future releases to further bridge the gap with datasets for other languages.

6. Conclusions

In this paper, we introduce an Arabic lipreading system that uses a fusion method of cross-attention mechanisms to decode spoken words from visual data. Our main contribution is to fuse facial landmark data with visual information, improve the accuracy of the system, and effectively address the limitations of lipreading data. Additionally, we present the larger-scale Arabic LRW-AR dataset to evaluate the system's performance. Experimental results demonstrate the effectiveness of our proposed system in predicting spoken words, achieving the best accuracy result of 85.85%. The proposed system represents a significant advancement in the field of visual Arabic speech recognition, highlighting the need for language-specific solutions. In our future work, we aim to extend our system to support sentence-level lipreading. These advancements will enable real-time application development and further improve continuous speech recognition, especially for the Arabic system.

Author Contributions: Conceptualization, S.D., A.B.-H. and A.R.; methodology, S.D., A.B.-H. and A.R.; software, S.D., A.B.-H. and A.R.; validation, A.B.-H.; formal analysis, A.R.; investigation, S.D.; data curation, S.D.; writing—original draft preparation, S.D.; writing—review and editing, A.B.-H., A.R. and A.K.; visualization, S.D.; supervision, A.B.-H. and A.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset generated during the current study is publicly available (https://osf.io/rz49x, accessed on 27 December 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ivanko, D.; Ryumin, D.; Karpov, A. Automatic Lip-Reading of Hearing Impaired People. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2019**, *42*, 97–101. [CrossRef]
2. Rothkrantz, L. Lip-reading by surveillance cameras. In Proceedings of the 2017 Smart City Symposium Prague (SCSP), Prague, Czech Republic, 25–26 May 2017; pp. 1–6.
3. Lu, L.; Yu, J.; Chen, Y.; Liu, H.; Zhu, Y.; Kong, L.; Li, M. Lip reading-based user authentication through acoustic sensing on smartphones. *IEEE/ACM Trans. Netw.* **2019**, *27*, 447–460. [CrossRef]
4. Chowdhury, D.P.; Kumari, R.; Bakshi, S.; Sahoo, M.N.; Das, A. Lip as biometric and beyond: A survey. *Multimed. Tools Appl.* **2022**, *81*, 3831–3865. [CrossRef]
5. Sheng, C.; Kuang, G.; Bai, L.; Hou, C.; Guo, Y.; Xu, X.; Pietikäinen, M.; Liu, L. Deep Learning for Visual Speech Analysis: A Survey. *arXiv* **2022**, arXiv:2205.10839. [CrossRef]
6. Ben-Hamadou, A. Control Method, Control Device, System and Motor Vehicle Comprising Such a Control Device. US Patent 10,627,898, 21 April 2020.
7. Dhanjal, A.S.; Singh, W. A comprehensive survey on automatic speech recognition using neural networks. *Multimed. Tools Appl.* **2024**, *83*, 23367–23412. [CrossRef]
8. Li, D.; Gao, Y.; Zhu, C.; Wang, Q.; Wang, R. Improving Speech Recognition Performance in Noisy Environments by Enhancing Lip Reading Accuracy. *Sensors* **2023**, *23*, 2053. [CrossRef]
9. Akhter, N.; Ali, M.; Hussain, L.; Shah, M.; Mahmood, T.; Ali, A.; Al-Fuqaha, A. Diverse Pose Lip-Reading Framework. *Appl. Sci.* **2022**, *12*, 9532. [CrossRef]
10. Fenghour, S.; Chen, D.; Guo, K.; Li, B.; Xiao, P. Deep learning-based automated lip-reading: A survey. *IEEE Access* **2021**, *9*, 121184–121205. [CrossRef]
11. Jitaru, A.C.; Ştefan, L.D.; Ionescu, B. Toward Language-independent Lip Reading: A Transfer Learning Approach. In Proceedings of the 2021 International Symposium on Signals, Circuits and Systems (ISSCS), Iasi, Romania, 15–16 July 2021; pp. 1–4.
12. Chung, J.S.; Zisserman, A. Lip Reading in the Wild. In *Computer Vision—ACCV 2016*; Lai, S.H., Lepetit, V., Nishino, K., Sato, Y., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2017; Volume 10112; pp. 87–103.
13. Yang, S.; Zhang, Y.; Feng, D.; Yang, M.; Wang, C.; Xiao, J.; Long, K.; Shan, S.; Chen, X. LRW-1000: A Naturally-Distributed Large-Scale Benchmark for Lip Reading in the Wild. In Proceedings of the 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), Lille, France, 14–18 May 2019; pp. 1–8.
14. Egorov, E.; Kostyumov, V.; Konyk, M.; Kolesnikov, S. LRWR: Large-Scale Benchmark for Lip Reading in Russian language. *arXiv* **2021**, arXiv:2109.06692.
15. Sheng, C.; Zhu, X.; Xu, H.; Pietikäinen, M.; Liu, L. Adaptive Semantic-Spatio-Temporal Graph Convolutional Network for Lip Reading. *IEEE Trans. Multimed.* **2021**, *24*, 3545–3557. [CrossRef]
16. Daou., S.; Rekik., A.; Ben-Hamadou., A.; Kallel., A. Near-infrared Lipreading System for Driver-Car Interaction. In Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, Lisbon, Portugal, 19–21 February 2023.
17. Cookea, M.; Barker, J.; Cunningham, S.; Shao, X. An audio-visual corpus for speech perception and automatic speech recognition. *J. Acoust. Soc. Am.* **2006**, *120*, 5. [CrossRef] [PubMed]
18. Nagrani, A.; Chung, J.S.; Zisserman, A. VoxCeleb: A Large-Scale Speaker Identification Dataset. *Telephony* **2017**, *3*, 33–039.
19. Chung, J.S.; Senior, A.W.; Vinyals, O.; Zisserman, A. Lip Reading Sentences in the Wild. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3444–3453.
20. Chung, J.S.; Zisserman, A. Lip Reading in Profile. In Proceedings of the British Machine Vision Conference, London, UK, 4–7 September 2017.
21. Afouras, T.; Chung, J.S.; Senior, A.; Vinyals, O.; Zisserman, A. Deep Audio-Visual Speech Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *44*, 8717–8727. [CrossRef] [PubMed]
22. Afouras, T.; Chung, J.S.; Zisserman, A. LRS3-TED: A large-scale dataset for visual speech recognition. *Corpus* **2008**, *123*, 3878.
23. Chung, J.S.; Nagrani, A.; Zisserman, A. VoxCeleb2: Deep Speaker Recognition. *arXiv* **2018**, arXiv:1806.05622.
24. Chung, J.S.; Nagrani, A.; Coto, E.; Xie, W.; McLaren, M.; Reynolds, D.A.; Zisserman, A. VoxSRC 2019: The first VoxCeleb Speaker Recognition Challenge. *arXiv* **2019**, arXiv:1912.02522.
25. Sagheer, S.A.A. Audio-Visual Arabic Speech (AVAS) Database for Human-Computer Interaction Applications. *Int. J. Adv. Res. Comput. Sci. Softw. Eng.* **2013**, *3*.

26. Dweik, W.; Altorman, S.; Ashour, S. Read my lips: Artificial intelligence word-level arabic lipreading system. *Egypt. Inform. J.* **2022**, *23*, 1–12. [CrossRef]
27. Alsulami, N.H.; Jamal, A.T.; Elrefaei, L.A. Deep Learning-Based Approach for Arabic Visual Speech Recognition. *Comput. Mater. Contin.* **2022**, *71*, 85–108.
28. Malhotra, R.; Singh, P. Recent advances in deep learning models: A systematic literature review. *Multimed. Tools Appl.* **2023**, *82*, 44977–45060. [CrossRef]
29. Gutierrez, A.; Robert, Z. Lip Reading Word Classification. Technical Report, Standford University, CS231 Project Report. 2017. Available online: https://cs231n.stanford.edu/reports/2017/pdfs/227.pdf (accessed on 23 December 2024).
30. Rekik, A.; Ben-Hamadou, A.; Mahdi, W. A new visual speech recognition approach for RGB-D cameras. In Proceedings of the International Conference on Image Analysis and Recognition, Vilamoura, Portugal, 22–24 October 2014; Springer: Cham, Switzerland, 2014; pp. 21–28.
31. Assael, Y.M.; Shillingford, B.; Whiteson, S.; De Freitas, N. LipNet: End-to-End Sentence-level Lipreading. *arXiv* **2016**, arXiv:1611.01599.
32. Petridis, S.; Stafylakis, T.; Ma, P.; Cai, F.; Tzimiropoulos, G.; Pantic, M. End-to-end audiovisual speech recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, Canada, 15–20 April 2018; pp. 6548–6552.
33. Stafylakis, T.; Tzimiropoulos, G. Combining Residual Networks with LSTMs for Lipreading. *arXiv* **2017**, arXiv:1703.04105.
34. Zhao, X.; Yang, S.; Shan, S.; Chen, X. Mutual information maximization for effective lip reading. In Proceedings of the 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), Buenos Aires, Argentina, 16–20 November 2020; pp. 420–427.
35. Ma, P.; Martinez, B.; Petridis, S.; Pantic, M. Towards Practical Lipreading with Distilled and Efficient Models. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, Canada, 6–11 June 2021; pp. 7608–7612.
36. Martinez, B.; Ma, P.; Petridis, S.; Pantic, M. Lipreading using Temporal Convolutional Networks. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6319–6323.
37. Koumparoulis, A.; Potamianos, G. Accurate and resource-efficient lipreading with efficientnetv2 and transformers. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 8467–8471.
38. Miled, M.; Messaoud, M.A.B.; Bouzid, A. Lip reading of words with lip segmentation and deep learning. *Multimed. Tools Appl.* **2023**, *82*, 551–571. [CrossRef]
39. Naderi, N.; Nasersharif, B. Cross Corpus Speech Emotion Recognition using transfer learning and attention-based fusion of Wav2Vec2 and prosody features. *Knowl.-Based Syst.* **2023**, *277*, 110814. [CrossRef]
40. Lohrenz, T.; Möller, B.; Li, Z.; Fingscheidt, T. Relaxed attention for transformer models. In Proceedings of the 2023 International Joint Conference on Neural Networks (IJCNN), Gold Coast, Australia, 18–23 June 2023; pp. 1–10.
41. Nagrani, A.; Albanie, S.; Zisserman, A. Seeing voices and hearing faces: Cross-modal biometric matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8427–8436.
42. Tsai, Y.H.H.; Bai, S.; Liang, P.P.; Kolter, J.Z.; Morency, L.P.; Salakhutdinov, R. Multimodal transformer for unaligned multimodal language sequences. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 27 July–2 August 2019; Volume 2019, p. 6558.
43. Rajan, V.; Brutti, A.; Cavallaro, A. Is cross-attention preferable to self-attention for multi-modal emotion recognition? In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 4693–4697.
44. Sagonas, C.; Tzimiropoulos, G.; Zafeiriou, S.; Pantic, M. 300 Faces in-the-Wild Challenge: The First Facial Landmark Localization Challenge. In Proceedings of the 2013 IEEE International Conference on Computer Vision Workshops, Sydney, Australia, 2–8 December 2013; pp. 397–403.
45. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
46. Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y. Graph attention networks. *arXiv* **2017**, arXiv:1710.10903.
47. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
48. King, D.E. Dlib-ml: A Machine Learning Toolkit. *J. Mach. Learn. Res.* **2009**, *10*, 1755–1758.
49. Lucas, B.D.; Kanade, T.; et al. An Iterative Image Registration Technique with an Application to Stereo Vision. In Proceedings of the 7th International Conference on Artificial Intelligence (IJCAI), Vancouver, Canada, 24–28 August 1981.

50. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 8026–8037.
51. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. *arXiv* **2017**, arXiv:1711.05101.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.