

GeoVisual-LipNet System : A Two-Stream 3D CNN–GRU Architecture with Lip Geometry

H. Benaliouche, D. Bellibel and M. Habbati

Keywords and phrases: lip reading, GRID dataset, visual speech recognition, phoneme mapping, visemes.

The abstract of the preliminary manuscript entitled ‘A Two-Stream 3D CNN–GRU Model for Lip Reading with Geometric Feature Integration’ was presented at ICRAMI 2025, Tunisia, in November 2025.

The authors would like to thank the reviewers and editor for their constructive comments and valuable suggestions that improved the quality of our paper.

Abstract Lip reading, or visual speech recognition, aims to recognize spoken content by analyzing the movements of a speaker’s lips. Although deep learning has improved performance in this field, visual-only approaches still struggle with ambiguous lip shapes, speaker variability, and sensitivity to lighting and texture. To address these issues, this work proposes a dual-stream lip reading framework that combines appearance-based visual features with explicit geometric information extracted from lip landmarks. The visual stream uses a 3D convolutional neural network followed by bidirectional GRUs to capture spatiotemporal patterns in lip image sequences, while the geometric stream models articulatory motion using landmark coordinates. The two streams are fused at an intermediate feature level and decoded with Connectionist Temporal Classification, enabling end-to-end training without frame-level alignment. Experiments on the GRID corpus show that integrating geometric motion cues with visual features improves recognition accuracy, achieving a word error rate of 1.64% and a character error rate of 0.63%, demonstrating the benefit of combining complementary visual and geometric representations.

1 Introduction

Lip reading, also known as visual speech recognition, focuses on understanding spoken language by observing the movements of a speaker’s lips [1]. This capability is especially valuable in situations where audio signals are unreliable or unavailable, such as in noisy environments, silent communication settings, assistive technologies for people with hearing impairments, or certain human–computer interaction scenarios.

Unlike traditional speech recognition, lip reading relies entirely on visual information. This makes the task inherently challenging. Many phonemes share similar lip shapes, which creates visual ambiguity, and the absence of acoustic cues makes it harder to distinguish between them. In addition, variations in speaker appearance, lighting conditions, camera quality, and recording environments can further complicate recognition.

Early lip reading systems were based on handcrafted visual features, including Histogram of Oriented Gradients (HOG), geometric lip descriptors, and optical flow [2] [3][4][5]. These features were typically combined with classical classifiers such as Support Vector Machines (SVMs) or Hidden Markov Models (HMMs). Although these approaches provided some initial success, they struggled to capture the complex spatiotemporal nature of speech and often failed to generalize well across different speakers.

The emergence of deep learning brought significant improvements. Data-driven methods us-

ing Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) became the dominant approach. Many systems employed 2D CNNs to extract spatial features from each frame, followed by recurrent layers to model temporal dependencies [6][7][8][9]. However, processing frames independently prevents these models from directly capturing short-term motion patterns between consecutive frames.

To overcome this limitation, researchers introduced 3D CNN-based architectures that jointly model spatial and temporal information. Systems such as LipNet [10] showed that combining 3D CNNs with recurrent networks and Connectionist Temporal Classification (CTC) enables effective end-to-end lip reading. Despite these advances, most deep learning approaches still rely entirely on raw visual appearance.

This dependence on pixel-based features introduces several drawbacks. Models can become sensitive to changes in lighting, texture, and background. Pooling operations may also discard fine geometric details, and appearance alone is often insufficient to distinguish between visually similar phonemes.

In contrast, explicit lip geometry provides additional articulatory information. Lip landmark coordinates describe the precise shape and motion of the mouth, including opening, closure, rounding, and stretching. These geometric features are compact, interpretable, and largely unaffected by variations in lighting or texture. Recent advances in deep learning have significantly improved visual recognition tasks across various domains [11], particularly through the integration of multiple feature representations. In this context, appearance-based and geometry-based representations offer complementary strengths. Visual features capture texture and shading patterns, while landmark trajectories encode the underlying articulatory motion. Combining both sources of information can therefore lead to more robust and discriminative representations for visual speech recognition.

Motivated by this idea, we propose a dual-stream architecture that processes visual appearance and lip landmark trajectories in parallel, and fuses them at the feature level before sequence decoding.

We summarize our main contributions as follows:

- (i) We introduce a unified lip reading model that processes appearance-based visual features and explicit lip landmark trajectories simultaneously, allowing texture and articulatory motion to be modeled together.
- (ii) Instead of relying on complex attention or multi-stage fusion strategies, the proposed approach uses an intermediate feature-level fusion mechanism that preserves complementary information while reducing computational complexity.
- (iii) The proposed design integrates both streams into a single end-to-end trainable system using Connectionist Temporal Classification (CTC) loss, eliminating the need for frame-level annotations and enabling implicit temporal alignment between visual inputs and target sequences.

2 Related Work

Lip reading, or visual speech recognition, aims to recognize spoken language using only visual information from mouth movements. This task is challenging due to the ambiguity of lip shapes, speaker variability, and the absence of acoustic cues. Over time, approaches have evolved from handcrafted feature-based systems to deep learning architectures [1].

2.1 Handcrafted Visual Features

Early lip reading systems relied on handcrafted visual descriptors extracted from the mouth region, such as geometric features, optical flow, or Histogram of Oriented Gradients (HOG) [2] [3][4][5]. These features were typically combined with classical classifiers such as Support Vector Machines or Hidden Markov Models. Although interpretable, such methods struggled to model the complex spatiotemporal dynamics of speech and were sensitive to illumination changes, camera quality, and speaker appearance. As a result, their performance was limited, especially for continuous speech recognition.

2.2 2D CNNs with Recurrent Models

With the rise of deep learning, 2D Convolutional Neural Networks were introduced to automatically learn visual features from lip images [6][7][8][9]. In these approaches, each frame was processed independently by a CNN, while temporal dependencies were modeled using recurrent networks such as LSTMs. Although these systems improved performance compared to hand-crafted methods, they treated time implicitly and could not directly capture short-term motion patterns across consecutive frames, limiting their ability to model subtle articulatory dynamics.

2.3 3D CNNs, End-to-End Learning, and CTC

To address the limitations of frame-based processing, 3D CNNs were introduced to capture spatiotemporal features directly from short video segments [12]. These models apply convolutions across both spatial and temporal dimensions, enabling the learning of motion-aware representations. Recurrent networks such as LSTMs and GRUs were then used to model long-range temporal dependencies.

A landmark work in this direction is LipNet [10], which combined a 3D CNN front-end with recurrent layers and Connectionist Temporal Classification (CTC) [13] for end-to-end sequence prediction without frame-level annotations. Subsequent studies explored the integration of explicit geometric information, such as facial landmarks or lip contours, to complement appearance-based features [14]. Other works investigated multimodal fusion and robustness improvements under challenging visual conditions [15][16][17][18].

Despite these advances, many deep models still rely primarily on appearance-based features and do not explicitly enforce lip geometry. This can lead to inconsistent motion representations across speakers, reduced robustness under pose or lighting variations, and confusion between visually similar phonemes.

Recent architectures further explored combinations of 3D CNNs, recurrent models, and CTC. Jeon *et al.* [19] proposed a sentence-level lipreading system based on multiple 3D CNN variants followed by bidirectional GRUs, achieving improved error rates on the GRID dataset. Similarly, Tabassum *et al.* [20] developed a visual-to-sound reconstruction framework using a 3D CNN and bidirectional GRUs with CTC, along with a lightweight coordinate-based model for real-time phrase recognition.

While these works confirm the effectiveness of spatiotemporal visual modeling, the proposed method differs by integrating appearance-based features and explicit lip landmark trajectories within a unified dual-stream architecture. By fusing visual and geometric representations at the feature level before CTC decoding, the proposed approach aims to preserve fine articulatory cues while improving robustness to illumination changes, blur, and speaker variability.

2.4 Key Limitation in Previous Methods

A key limitation observed in many previous deep lip reading models is that they do not explicitly enforce or preserve **lip geometry**. Instead, these approaches rely primarily on CNN-based feature extraction to infer geometric structure indirectly from raw pixel intensities. While this strategy enables the learning of powerful appearance-based representations, it does not guarantee that fine-grained articulatory cues such as precise lip contour changes, closure duration, or rounding dynamics are consistently encoded in the learned features.

This limitation becomes particularly evident when considering the concept of **visemes**, which are visually distinguishable mouth shapes corresponding to groups of phonemes. As illustrated in Figure 1, multiple phonemes often share highly similar visual appearances, and their discrimination depends on subtle geometric differences and temporal evolution rather than static texture information alone.

As a consequence, the absence of explicit geometric modeling can lead to several practical issues:

- **Inconsistent motion representation across speakers:** variations in mouth shape, facial structure, skin texture, illumination conditions, or facial hair can cause identical articulatory motions to be represented differently in pixel space, reducing cross-speaker generalization.

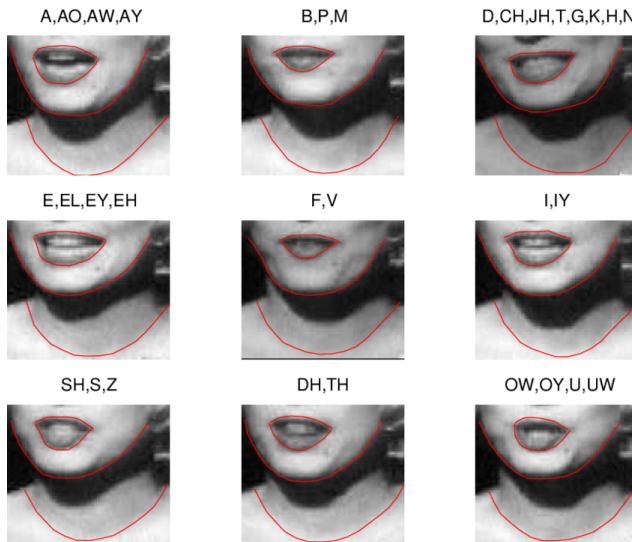


Figure 1. Characteristic viseme examples illustrating groups of phonemes with similar visual mouth appearances. Despite different acoustic realizations, many phonemes share comparable lip shapes, highlighting the importance of explicit geometric modeling for reliable visual speech recognition.

- **Reduced robustness under slight pose variations:** small head movements or changes in camera viewpoint may significantly alter pixel-level appearance, even when the underlying lip geometry and articulation remain unchanged.
- **Confusion between visually similar phonemes:** many phoneme groups (e.g., bilabial or rounded sounds) produce nearly identical mouth appearances in isolated frames. Accurate discrimination in such cases relies heavily on subtle geometric cues and temporal motion patterns, which may be partially lost through convolution and pooling operations.

These observations motivate the incorporation of an explicit geometric representation, such as lip landmark trajectories, to provide stable articulatory constraints that complement appearance-based visual features. By modeling lip geometry directly, the system can better preserve motion consistency, improve robustness to speaker and pose variations, and reduce ambiguity between visually similar phonemes.

3 Proposed Method

To address the lack of explicit geometric preservation in prior approaches, we propose **GeoVisual-LipNet**, a dual-stream lip reading framework that integrates learned visual representations with explicit lip landmark motion.

3.1 Overview of GeoVisual-LipNet

GeoVisual-LipNet is designed as an end-to-end trainable architecture that processes appearance-based and geometry-based information in parallel, allowing the model to capture both visual texture cues and precise articulatory dynamics.

The proposed model is trained and evaluated on the GRID audio-visual speech corpus[21], a widely used benchmark dataset for visual speech recognition. GRID provides fixed-length utterances recorded from multiple speakers under controlled conditions, making it suitable for evaluating both recognition accuracy and speaker generalization. By applying GeoVisual-LipNet to this dataset, we assess the effectiveness of combining appearance-based and geometry-based representations for lip reading.

- **Parallel geometric stream using explicit landmarks:** we introduce a dedicated stream

that processes lip landmark coordinates extracted for each frame, providing an explicit representation of mouth geometry and reducing reliance on pixel-only inference.

- **BiGRU modeling of articulatory dynamics:** a bidirectional GRU (BiGRU) encodes the temporal evolution of landmark trajectories, capturing motion patterns such as closure duration, opening rate, and rounding dynamics that are critical for distinguishing visually similar phonemes. GRUs are used because they provide efficient temporal modeling while using fewer parameters than LSTMs. A GRU relies on two gates: the update gate, which controls how much past information is preserved, and the reset gate, which determines how much of the previous state is ignored when computing new features. This simpler structure reduces computational cost and training time while maintaining strong sequence modeling capability.
- **Intermediate fusion before CTC classification:** visual features (from the 3D CNN + BiGRU stream) and geometric features (from the landmark BiGRU stream) are fused at the feature level prior to the final CTC-based classifier. This allows the model to combine appearance cues with geometric motion constraints, improving robustness across speakers and under minor pose variations.

The bidirectional formulation enables the network to exploit both past and future context at each time step, which is particularly important in lip reading where individual frames are often ambiguous without surrounding motion information. This balance between modeling capacity and computational efficiency makes BiGRUs well suited for visual speech recognition tasks.

3.2 Dataset: GRID Corpus

The **GRID dataset** is a large-scale audio-visual speech corpus widely used for research in automatic speech recognition and lip reading. It contains recordings from **34 speakers (18 male and 16 female)**, where each speaker utters **1,000 short sentences**, resulting in a total of **34,000 audio-visual utterances** [21]. Each utterance is recorded in controlled conditions with a frontal view of the speaker’s face, making the corpus suitable for extracting mouth regions and studying visual speech dynamics.

Sentence Structure and Grammar

A key characteristic of GRID is its **fixed grammar**, where each sentence follows the same template, which simplifies analysis and enables controlled evaluation. An example of fixed sentence grammar structure is illustrated in Figure 2.

command color preposition letter digit adverb

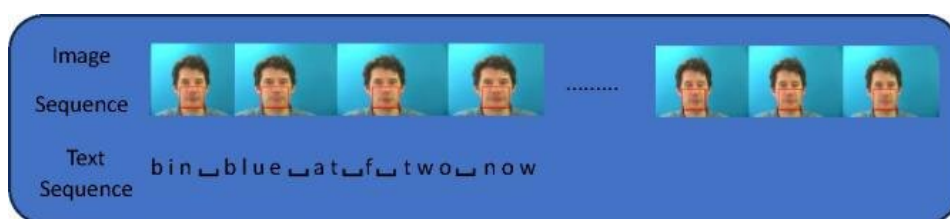


Figure 2. Fixed sentence grammar structure in the GRID corpus.

File Naming Convention

Each sentence is also associated with a short file identifier. For example, the sentence “bin blue at two now” can be represented by a name such as:

bba2n

This identifier encodes the main tokens of the sentence in a compact form, and it is used consistently across the video, audio, and annotation files.

Utterance Properties and Alignments

Each GRID utterance has the following properties:

- **Duration:** approximately **3 seconds** per sentence.
- **Recording quality:** high-quality **video and audio** captured under consistent conditions.
- **Alignment files:** each utterance includes a time-stamped **.align** file that provides word-level segmentation boundaries.

The **.align** files are particularly important for training and evaluation, as they provide the ground-truth transcript and temporal segmentation information without requiring manual frame-level labeling. In our pipeline, these files are used to build the target text sequences for CTC training while allowing the model to learn alignments automatically.

3.3 Data Preparation and Preprocessing

Data preparation consists of extracting visual frames and lip landmark coordinates from each video, followed by normalization and sequence standardization. These steps ensure that both visual and geometric inputs are temporally aligned and suitable for batch-based training.

Frame Extraction

Each utterance in the GRID dataset is stored as an MPEG (.mpg) video file with a duration of approximately three seconds. The first step of the GeoVisual-LipNet system consists of extracting individual frames from these recordings.

For each video:

- the video file is opened using a decoding library,
- frames are read sequentially from start to end,
- each frame is saved as a .jpg image,
- frames are indexed to preserve temporal order.

Extracted frames are organized in a hierarchical directory structure based on speaker and video identifiers.

Each GRID utterance contains **75 frames** at a frame rate of 25 fps, providing a consistent temporal length that simplifies sequence modeling. Figure 3 illustrates an example of a single video frame extracted from the GRID dataset.



Figure 3. Example of a single video frame extracted from the GRID dataset, illustrating the mouth region used as input to the visual stream.

Lip Landmark Extraction

In parallel with frame extraction, explicit lip geometry is obtained using facial landmark detection. For each frame:

- facial landmarks are detected using the **dlib** library,
- the standard 68-point landmark model is applied,
- only mouth landmarks (indices 48–67) are retained.

Each frame is therefore represented by:

$$20 \text{ lip landmarks} \times (x, y) = 40 \text{ values.}$$

These coordinates describe the geometric structure of the lips and capture articulatory motion such as opening, closure, rounding, and stretching. Unlike raw pixel values, landmark coordinates are largely invariant to illumination, color, and texture variations.

For each frame, the selected landmarks are stored as (x, y) coordinate pairs in JSON format, forming a temporal trajectory of lip motion across the utterance. Figure 4 illustrates the facial landmark detection process.

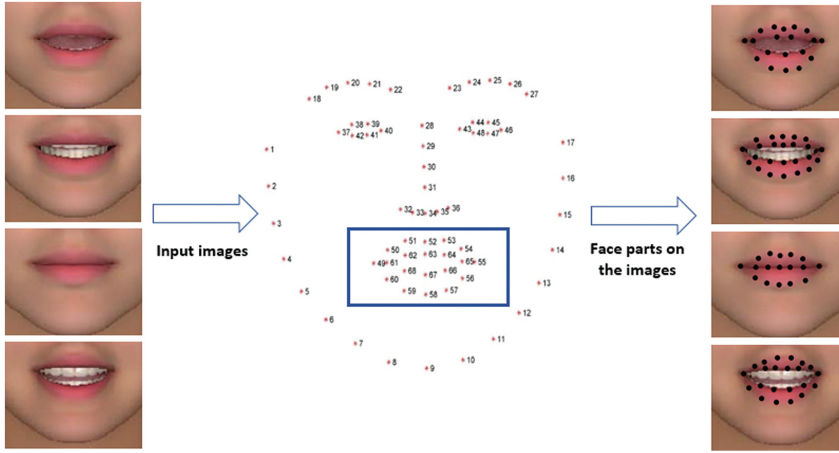


Figure 4. Facial landmark detection using dlib, showing the full landmark set and the subset corresponding to the lip region.

Normalization and Sequence Preparation

Preprocessing is applied separately to visual frames and landmark coordinates to standardize inputs and enable stable training.

Frame preprocessing. Each frame is resized to a fixed resolution of 64×128 pixels. Pixel values are normalized to the range $[0, 1]$ by dividing by 255. During training, horizontal flipping is applied as data augmentation to improve robustness to pose variations.

Sequences are padded temporally to a fixed length to allow batch processing.

Coordinate preprocessing. For each frame, the $(20, 2)$ landmark coordinates are:

- Normalized by image width and height,
- Flattened into a 40-dimensional vector,
- Temporally padded to match the visual sequence length.

Input Tensor Shapes

After preprocessing, the inputs to the model are represented as:

Visual stream input:

$$x \in \mathbb{R}^{(B,3,T,64,128)},$$

where B is the batch size and T is the temporal length.

Geometric stream input:

$$c \in \mathbb{R}^{(T,B,40)},$$

where each time step contains the normalized lip landmark coordinates.

3.4 System Architecture

The proposed **GeoVisual-LipNet system** is a dual-stream, end-to-end trainable architecture designed to jointly model visual appearance and explicit lip geometry for robust lip reading. The system processes two complementary sources of information in parallel: RGB mouth image sequences and lip landmark trajectories. These streams are temporally aligned and fused at an intermediate representation level before sequence decoding. Figure 5 presents an overview of the GeoVisual-LipNet system architecture.

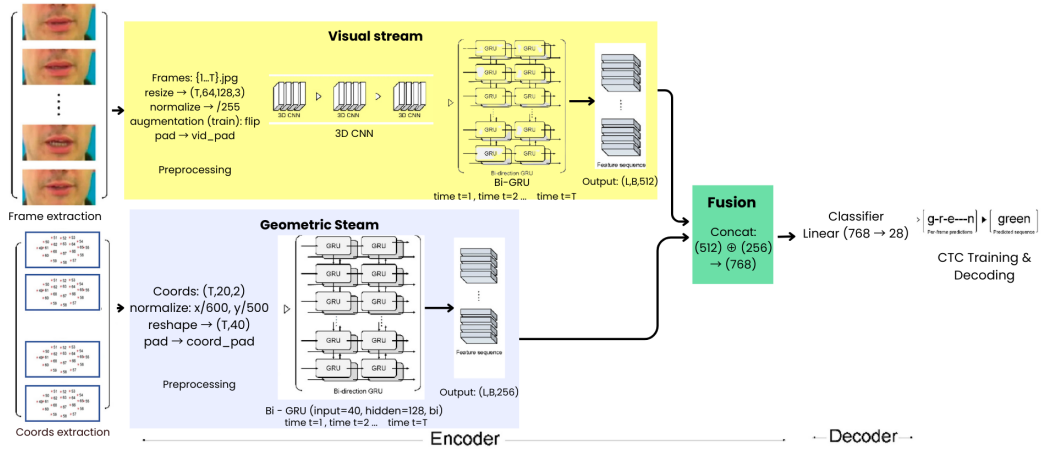


Figure 5. Overview of the GeoVisual-LipNet system architecture. The model consists of a visual stream based on a 3D CNN and Bi-GRU, a geometric stream based on lip landmark trajectories and a Bi-GRU, followed by intermediate feature fusion and CTC-based decoding.

Visual Stream

The visual stream learns appearance-based speech information directly from lip image sequences. Its objective is to capture both the spatial structure of the mouth and the short-term motion patterns that correspond to phonetic articulation.

After preprocessing, each video is represented as:

$$x \in \mathbb{R}^{(B,3,T,64,128)},$$

where B is the batch size, T is the number of frames, and 64×128 is the spatial resolution of the cropped mouth region.

3D CNN feature extraction. A 3D Convolutional Neural Network is used as the front-end of the visual stream. Unlike a 2D CNN, which processes each frame independently, the 3D CNN performs convolution across both spatial and temporal dimensions. This enables the model to capture short-term motion cues such as lip closure, opening dynamics, and transitions between mouth shapes.

After three convolution and pooling stages, the output tensor becomes:

$$x_{\text{cnn}} \in \mathbb{R}^{(B,96,T,4,8)}.$$

Temporal modeling with Bi-GRU. The CNN output is reshaped into a temporal sequence:

$$x_{\text{seq}} \in \mathbb{R}^{(T,B,96 \cdot 4 \cdot 8)}.$$

Two stacked bidirectional GRU layers are then applied to model long-range temporal dependencies. The bidirectional design enables the network to use both past and future context at each time step.

The final output of the visual stream is:

$$x_{\text{vis}} \in \mathbb{R}^{(T,B,512)}.$$

Geometric Stream

The geometric stream explicitly models lip motion using landmark trajectories. While the visual stream captures appearance information, this stream focuses on precise articulatory dynamics derived from lip geometry.

For each frame, 20 lip landmarks are extracted, resulting in a 40-dimensional vector representing the (x, y) coordinates of the mouth contour. After normalization and padding, the coordinate sequence is:

$$c_{\text{seq}} \in \mathbb{R}^{(T,B,40)}.$$

Since landmark coordinates are already compact and structured, only temporal modeling is required. A bidirectional GRU processes the coordinate sequence and learns motion patterns such as lip closure duration, opening speed, and transitions between rounded and stretched shapes.

The output of the geometric stream is:

$$x_{\text{geo}} \in \mathbb{R}^{(T,B,256)}.$$

Feature Fusion

After independent processing, the outputs of the visual and geometric streams are fused at the feature level (intermediate fusion). At this stage, both streams provide temporally contextualized representations.

- Visual features:

$$x_{\text{vis}} \in \mathbb{R}^{(T,B,512)}$$

- Geometric features:

$$x_{\text{geo}} \in \mathbb{R}^{(T,B,256)}$$

The two representations are concatenated along the feature dimension:

$$x_{\text{fused}} = x_{\text{vis}} \oplus x_{\text{geo}} \in \mathbb{R}^{(T,B,768)}.$$

This intermediate fusion strategy allows the model to combine appearance-based cues with explicit articulatory motion constraints, improving robustness to speaker variability and visually ambiguous phonemes.

CTC Classifier

The fused representation is passed through a linear classifier applied at each time step:

$$768 \rightarrow 28,$$

where 28 corresponds to the character vocabulary plus the CTC blank symbol.

This produces frame-level logits:

$$y \in \mathbb{R}^{(T,B,28)}.$$

Since explicit frame-to-character alignment is not available, the system is trained using **Connectionist Temporal Classification (CTC)**. CTC introduces a blank symbol and computes the probability of all valid alignments between the predicted frame-level outputs and the target transcription.

CTC provides several advantages for lip reading:

- Alignment-free training without frame-level annotations,
- Automatic temporal alignment between video frames and text,
- Support for variable-length input and output sequences,
- Robustness to repeated characters and silent frames.

These properties make CTC particularly suitable for visual speech recognition, where the timing of character boundaries is ambiguous and difficult to annotate manually.

3.5 Training Protocol

This section describes the experimental setup used to train and evaluate the proposed GeoVisual-LipNet model, including the dataset split, training hyperparameters, and implementation details.

Speaker-Independent Split

To evaluate the generalization capability of the model, the GRID dataset was divided at the speaker level. This prevents identity leakage between training and evaluation sets and ensures a realistic lip-reading scenario.

- 14 speakers for training
- 5 speakers for validation
- 5 speakers for testing

This speaker-independent protocol ensures that the model is evaluated on completely unseen identities, providing a fair assessment of its ability to generalize across speakers.

Training Hyperparameters

The model was trained end-to-end using parameters selected to ensure stable optimization and feasible GPU memory usage.

- **Optimizer:** Adam
- **Loss function:** Connectionist Temporal Classification (CTC)
- **Batch size:** 16
- **Learning rate:** 2×10^{-5}
- **Number of epochs:** 200

During training, both visual frames and lip landmark coordinates were processed simultaneously by the dual-stream architecture. Predictions were periodically decoded using greedy CTC decoding, and Character Error Rate (CER) and Word Error Rate (WER) were computed to monitor performance.

Validation was conducted at regular intervals using the dedicated validation speaker set. The same decoding strategy and evaluation metrics were applied. Final evaluation was performed on the held-out test speaker set without gradient computation.

3.6 Implementation Details

This section summarizes the practical configuration used to train the proposed architecture.

Hardware Configuration All experiments were conducted on a local workstation with the following specifications:

- **GPU:** NVIDIA GeForce RTX 4070
- **RAM:** 36 GB
- **CPU:** AMD Ryzen (8 cores)

Training was performed for up to 200 epochs. The average training time per epoch was approximately 120 minutes, resulting in a total training duration of roughly 10–15 days. This cost is mainly due to the large number of video sequences and the computational expense of processing both spatiotemporal video volumes and landmark trajectories.

Algorithm 1 Forward Pass of the Proposed Dual-Stream Lip Reading Model

```

procedure FORWARD( $X, C$ )                                ▷  $X$ : RGB frames,  $C$ : lip coordinates
  Input visual frames:  $X \in \mathbb{R}^{(B,3,T,64,128)}$ 
  Input coordinates:  $C \in \mathbb{R}^{(B,T,20,2)}$ 

   $X \leftarrow \text{Conv3D}_1(X)$                                 ▷ Visual Stream 3D CNN Feature Extraction
   $X \leftarrow \text{ReLU}(X)$ 
   $X \leftarrow \text{MaxPool3D}(X)$ 
   $X \leftarrow \text{Dropout3D}(X)$ 
   $X \leftarrow \text{Conv3D}_2(X)$ 
   $X \leftarrow \text{ReLU}(X)$ 
   $X \leftarrow \text{MaxPool3D}(X)$ 
   $X \leftarrow \text{Dropout3D}(X)$ 
   $X \leftarrow \text{Conv3D}_3(X)$ 
   $X \leftarrow \text{ReLU}(X)$ 
   $X \leftarrow \text{MaxPool3D}(X)$ 
   $X \leftarrow \text{Dropout3D}(X)$ 
   $X \leftarrow \text{Reshape}(X)$                                 ▷  $(B, 96, T, 4, 8) \rightarrow (T, B, F)$ 
                                                         ▷ Visual Stream Temporal Modeling

   $X \leftarrow \text{BiGRU}_1(X)$ 
   $X \leftarrow \text{Dropout}(X)$ 
   $X \leftarrow \text{BiGRU}_2(X)$ 
   $X \leftarrow \text{Dropout}(X)$ 

   $C \leftarrow \text{Reshape}(C)$                                 ▷ Geometric Stream Coordinate Processing
   $C \leftarrow \text{BiGRU}_{coord}(C)$                             ▷  $(B, T, 20, 2) \rightarrow (T, B, 40)$ 
   $C \leftarrow \text{Dropout}(C)$ 

   $F \leftarrow \text{Concatenate}(X, C)$                             ▷ Feature Fusion
                                                         ▷ Feature-level fusion
                                                         ▷ Classification
   $Y \leftarrow \text{Linear}(F)$                                 ▷ Per-frame character logits
  return  $Y$ 
end procedure

```

Framework and Model Implementation The proposed architecture was implemented using the PyTorch deep learning framework, which provides efficient support for tensor-based computation, automatic differentiation, and GPU acceleration. PyTorch was selected due to its flexibility in designing custom neural network modules and its suitability for sequence modeling tasks that combine convolutional and recurrent components.

GeoVisual-LipNet is designed as an end-to-end trainable dual-stream architecture that processes two complementary modalities: visual appearance and explicit lip geometry. The model is implemented as a unified computational graph in which both streams are trained jointly using a single loss function, allowing the network to learn consistent representations across modalities.

The first branch, referred to as the **visual stream**, operates directly on sequences of RGB lip images. This stream uses a 3D convolutional front-end to extract spatiotemporal features that jointly encode spatial appearance and short-term motion patterns. Unlike frame-based processing, the 3D convolutions allow the model to capture dynamic lip movements across consecutive frames. The resulting features are then passed through stacked bidirectional GRU layers, which model long-range temporal dependencies and incorporate contextual information from both past and future frames.

The second branch, referred to as the **geometric stream**, processes sequences of lip landmark coordinates. These coordinates explicitly describe the shape and motion of the lips over time and provide a compact representation that is largely invariant to illumination, texture, and speaker-specific appearance. Since the coordinate sequences are already low-dimensional and structured, this stream relies solely on a bidirectional GRU to model articulatory dynamics such as lip closure duration, opening speed, and shape transitions.

After independent processing, the outputs of the visual and geometric streams are fused at the feature level. This intermediate fusion allows the model to combine appearance-based cues with explicit motion constraints before making any classification decisions. The fused features are then passed through a linear layer that produces frame-level character logits.

Training is performed using Connectionist Temporal Classification (CTC) loss, which enables alignment-free sequence learning. This allows the model to convert variable-length visual inputs into text without requiring explicit frame-to-character annotations.

Algorithm 1 summarizes the forward computation of the proposed model.

The complete source code for data preprocessing, model training, and inference is available online at the following [link](#).

4 Performance Evaluation Metrics

The performance of the proposed lip reading system was evaluated using standard sequence recognition metrics commonly adopted in automatic speech recognition and visual speech recognition tasks. These metrics assess both the learning behavior during training and the final transcription accuracy of the model.

4.1 Learning Loss

During training, the learning progress of the model was monitored using the Connectionist Temporal Classification (CTC) loss. The training and validation losses provide insight into the convergence behavior and the generalization capability of the model. A decreasing loss indicates improved alignment between predicted character sequences and ground-truth transcriptions, while the gap between training and validation losses reflects the degree of overfitting.

4.2 Character Error Rate (CER) and Word Error Rate (WER)

To quantitatively assess transcription accuracy, we employed **Character Error Rate (CER)** and **Word Error Rate (WER)**. These metrics measure the discrepancy between the decoded output sequence and the ground-truth transcription based on edit distance.

Let N_C denote the total number of characters in the reference transcription and N_W the total number of words. Let S , D , and I represent the number of substitutions, deletions, and insertions, respectively.

The character error rate is defined as:

$$\text{CER} = \frac{C_S + C_D + C_I}{N_C} \times 100 \quad (4.1)$$

Similarly, the word error rate is defined as:

$$\text{WER} = \frac{W_S + W_D + W_I}{N_W} \times 100 \quad (4.2)$$

where the subscripts C and W denote character-level and word-level operations, respectively.

CER focuses on fine-grained character-level errors, while WER provides a higher-level evaluation of sentence correctness. Both metrics are necessary to fully characterize the performance of visual speech recognition systems.

4.3 Accuracy Metrics

In addition to error rates, we report character accuracy and word accuracy, which are derived directly from CER and WER.

$$\text{Character Accuracy} = 1 - \text{CER} \tag{4.3}$$

$$\text{Word Accuracy} = 1 - \text{WER} \tag{4.4}$$

These metrics provide an intuitive interpretation of recognition performance, where higher values indicate better transcription quality.

4.4 Decoding Strategy

During evaluation, frame-level character probabilities produced by the network were decoded using CTC-based decoding. The decoded sequence was obtained by selecting the most probable label at each time step, collapsing repeated characters, and removing blank symbols. This decoding strategy allows the model to generate variable-length transcriptions without explicit frame-level alignment.

5 Experimental Results

5.1 Training Behavior

Figure 6 illustrates the evolution of Character Error Rate (CER) and Word Error Rate (WER) during training over 200 epochs. Both metrics show a clear decreasing trend, indicating stable convergence of the model. Early epochs exhibit higher fluctuations due to the difficulty of learning temporal alignments with CTC, while later epochs show reduced variance as the model converges.

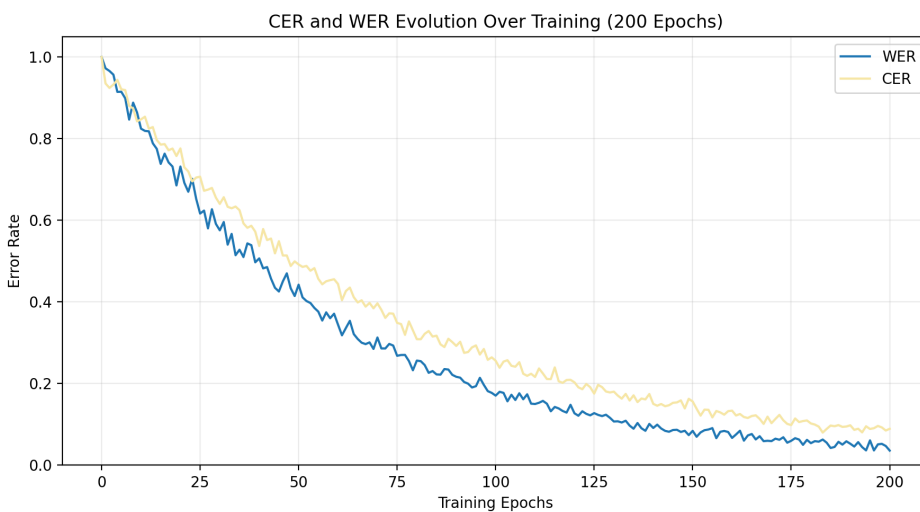


Figure 6. Evolution of Character Error Rate (CER) and Word Error Rate (WER) during training. Both curves decrease steadily with realistic fluctuations, reflecting stable CTC-based learning.

5.2 Quantitative Evaluation

Table 1 summarizes the final performance of the proposed model. The low CER and WER values demonstrate accurate character-level and word-level recognition. Correspondingly, high character and word accuracies confirm the robustness of the model across utterances.

Table 1. Quantitative performance of the proposed model on the GRID dataset

Metric	Value
Word Error Rate (WER)	0.0164
Character Error Rate (CER)	0.0063
Word Accuracy	0.9836
Character Accuracy	0.9937

The results indicate that the model correctly recognizes more than 98% of words and 99% of characters. The slightly higher WER compared to CER is expected, as a single character error may invalidate an entire word.

5.3 Qualitative Results

Figure 7 shows representative qualitative results comparing predicted transcriptions with ground-truth sentences. The examples demonstrate that the model accurately captures the grammatical structure and lexical content of GRID sentences, with only occasional minor errors.

predict	truth
BIN BLUE AT E FOUR NOW	BIN BLUE AT E FOUR NOW
BIN BLUE AT E FIVE SOON	BIN BLUE AT E FIVE SOON
BIN BLUE AT E SIX PLEASE	BIN BLUE AT E SIX PLEASE
BIN BLUE AN E SEVEN AGAIN	BIN BLUE AT E SEVEN AGAIN
BIN BLUE AT K EIGHT NOW	BIN BLUE AT K EIGHT NOW
BIN BLUE AT K NINE SOON	BIN BLUE AT K NINE SOON
BIN BLUE AT L ONE AGAIN	BIN BLUE AT L ONE AGAIN
BIN BLUE AN S ZERO PLEASE	BIN BLUE AT L ZERO PLEASE

test_iter=0, eta=0.4830, wer=0.0625, cer=0.0150, word_acc=0.9375, char_acc=0.9850

Figure 7. Qualitative comparison between predicted sentences (left) and ground-truth transcriptions (right). The model correctly reconstructs sentence structure and content with high accuracy.

Despite some errors, such as predicting "N" instead of "T" in "AT" (they have similar visual mouth appearances, as discussed in Section 4). These qualitative observations support the quantitative findings, confirming that the proposed fusion of visual and geometric streams enables reliable and context-aware lip reading. It is important to note that the incorporation of explicit lip landmark coordinates does not entirely eliminate all sources of visual ambiguity. In particular, certain phonemes such as "n" and "t" exhibit very similar viseme appearances, and occasional confusions between them were observed during real-time inference. This behavior is expected and reflects an inherent limitation of visual-only speech recognition, where multiple phonemes may share nearly identical articulatory configurations.

The primary role of the geometric coordinate stream is not to resolve all phoneme-level ambiguities, but rather to mitigate limitations associated with frame-based appearance modeling, such as sensitivity to illumination changes, motion blur, compression artifacts, and texture variability. By explicitly encoding lip geometry and motion trajectories, the coordinate stream stabilizes temporal representations and improves robustness across varying visual conditions. Consequently, occasional confusions between visually similar phonemes remain normal, even in the presence of geometric features, and do not indicate a failure of the proposed approach. A detailed error analysis reveals that most misclassifications occur between phonemes with highly similar viseme representations, highlighting the inherent ambiguity of visual-only speech signals.

5.4 Comparison with Existing Methods

To evaluate the effectiveness of the proposed approach, we compare its performance with representative baseline methods, the LipNet architecture, and more recent deep learning models on the GRID dataset. The comparison is reported using Character Error Rate (CER) and Word Error Rate (WER) under the unseen-speaker and overlapped-speaker protocols when available.

Table 2. Performance comparison on the GRID dataset (CER and WER in %).

Method	Unseen Speakers		Overlapped Speakers	
	CER	WER	CER	WER
Hearing-Impaired Person (avg) [10]	–	47.7	–	–
Baseline-LSTM [10]	38.4	52.8	15.2	26.3
Baseline-2D CNN [10]	16.2	26.7	4.3	11.6
Baseline-NoLM [10]	6.7	13.6	2.0	5.6
LipNet [10]	6.4	11.4	1.9	4.8
LCSNet [22]	1.0	2.3	–	–
Transformer-based model [23]	–	54.19	–	–
Proposed Method (3D CNN + Bi-GRU + coords)	4.1	7.3	0.6	1.6

The Transformer-based model in [23] reports a word-level accuracy of 45.81% on the GRID corpus rather than CER and WER, and therefore is not directly comparable within the same metric framework.

As shown in Table 2, the proposed model improves significantly over the original baselines and the LipNet architecture. Compared to LipNet, the proposed approach reduces the unseen-speaker CER from 6.4% to 3.1% and the WER from 11.4% to 6.3%.

Recent models such as LCSNet [22] achieve very low error rates by incorporating channel attention and selective feature fusion mechanisms. Although LCSNet reports lower CER and WER, it relies on more complex feature selection modules, whereas the proposed approach focuses on a dual-stream design that explicitly integrates geometric lip motion with appearance-based features.

The observed gains of the proposed model over classical baselines are explained by the dual-stream architecture. The visual stream (3D CNN + Bi-GRU) captures spatiotemporal appearance cues, while the geometric stream models explicit lip landmark motion. Fusing both representations at the feature level provides complementary constraints that reduce ambiguity in visually similar phonemes and improves robustness under illumination changes, blur, and appearance variations.

Overall, these results demonstrate that incorporating explicit geometric information alongside learned visual features leads to more discriminative and robust representations, improving recognition accuracy in both speaker-dependent and speaker-independent settings.

5.5 Inference Results

To assess how the proposed model behaves during inference, we conducted both offline video-based experiments and real-time tests. These experiments were designed to evaluate not only the recognition accuracy but also the stability of the system under different operating conditions.

Real-Time Video Inference

For the offline evaluation, a complete video sequence was processed using the trained model. Since full video clips cannot be embedded directly in the PDF, a representative frame is shown instead. During inference, the predicted transcription was drawn directly onto each frame, and the annotated frames were saved. This approach allowed the model’s predictions to be visually integrated into the video stream.

Figure 8 presents an example frame taken from the inference video. The text displayed on the frame corresponds to the model’s prediction at that moment. In this particular sequence, the predicted sentence matched the ground-truth transcription for the entire video.

The resulting offline inference performance was:

- Character Accuracy: 99%
- Word Accuracy: 98%

These results indicate that, when the model has access to the full temporal context of the sequence, it is able to reconstruct the target sentence with very high accuracy.

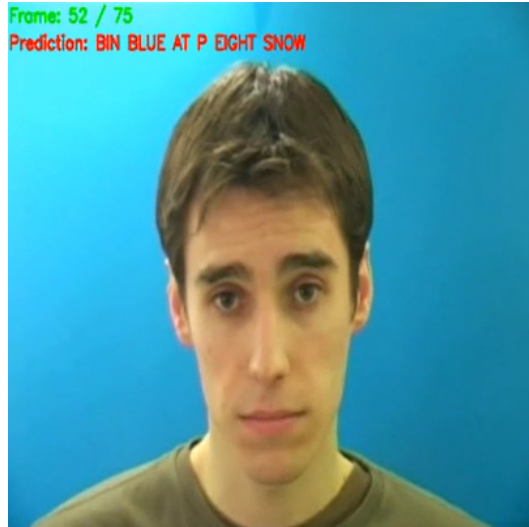


Figure 8. Example frame from the offline inference video with the predicted text overlaid on the image. The predicted sentence matches the ground truth.

Real-Time Inference Analysis

In addition to offline testing, the model was evaluated in a real-time scenario using live video input. Figure 8 illustrates an example frame from the offline inference video with the predicted text overlaid on the image. Overall, the system remained stable and produced consistent predictions. However, a small number of errors were observed, particularly cases where the character *e* was confused with *m*. This type of confusion is expected, as these characters can produce similar lip shapes, especially during fast articulation or when only limited temporal context is available.

Unlike offline inference, real-time processing does not always benefit from the full sequence context. This can slightly reduce the effectiveness of bidirectional temporal modeling, since the model may not have access to future frames when making predictions. Despite this limitation, such errors were infrequent and did not significantly affect the overall performance.

These observations suggest that the proposed system is robust in both offline and real-time conditions, while also highlighting areas for improvement. Future work could incorporate temporal smoothing techniques or language-model-based decoding to further stabilize predictions during real-time operation.

6 Conclusion and Future Work

This work introduced a dual-stream visual speech recognition system that combines appearance-based features from video frames with explicit geometric information derived from lip landmark coordinates. The visual stream uses a 3D convolutional neural network to capture short-term spatiotemporal motion, while bidirectional GRUs model long-range temporal dependencies across the sequence.

The addition of a geometric stream provides explicit articulatory cues that are not always preserved in pixel-based representations. Rather than replacing visual features, the landmark-based stream complements them by offering greater stability under variations in lighting, blur,

and texture. By fusing both streams at the feature level, the model is able to combine motion-aware geometric information with learned visual representations before decoding.

Experimental results on the GRID dataset demonstrate strong recognition performance, including near-perfect character- and word-level accuracy during offline inference. Minor errors observed during real-time testing, such as confusion between visually similar phonemes, are consistent with known limitations of visual-only speech recognition. The coordinate stream was not intended to eliminate all viseme ambiguities, but rather to stabilize appearance-based modeling and reduce sensitivity to visual variations.

Overall, the proposed architecture shows that combining appearance and geometric representations leads to a more robust lip-reading system. The framework is flexible and can be extended in several directions to further improve performance, particularly in real-time scenarios.

Several avenues for future work can be explored:

- **Enhanced Data Augmentation:** Introduce temporally consistent augmentations such as random frame dropping, temporal jittering, variable playback speed, and simulated motion blur. These strategies may help the model generalize better to real-time video streams, where frame quality and timing are less predictable.
- **Speaker and Pose Variability:** Extend the training data with synthetic pose variations, mild head rotations, and illumination changes to improve robustness under real-world recording conditions.
- **Language Model Integration:** Combine the CTC decoder with an external character-level or word-level language model to reduce linguistically invalid predictions and improve decoding stability during real-time inference.
- **Evaluating the proposed dual-stream architecture on other large-scale datasets:** Future work will extend the evaluation of the proposed model to larger and more challenging datasets such as LRW [24] and LRS2 [25], in order to further assess its generalization capability under unconstrained conditions.

References

- [1] Z. Zhou, G. Zhao, X. Hong, and M. Pietikäinen, *A review of recent advances in visual speech decoding*, Image and Vision Computing, vol. 32, no. 9, pp. 590–605, (2014). doi: 10.1016/j.imavis.2014.06.004.
- [2] I. Matthews, T. Cootes, and J. Bangham, *Extraction of visual features for lipreading*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 2, pp. 198–213, (2002), doi: 10.1109/34.982900.
- [3] J. N. Gowdy, A. Subramanya, C. Bartels, and J. Bilmes, *DBN based multi-stream models for audio-visual speech recognition*, in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), vol. 1, pp. I–993, (2004).
- [4] K. Livescu, O. Cetin, M. Hasegawa-Johnson, S. King, C. Bartels, N. Borges, A. Kantor, P. Lal, L. Yung, A. Bezman, et al., *Articulatory feature-based methods for acoustic and audio-visual speech recognition*, in Proc. IEEE ICASSP, vol. 4, pp. IV–621–IV–624, (2007). doi: 10.1109/ICASSP.2007.367248.
- [5] E.-J. Ong and R. Bowden, *Learning temporal signatures for lip reading*, in Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops), pp. 958–965, (2011). doi: 10.1109/ICCVW.2011.6130427.
- [6] J. S. Chung and A. Zisserman, *Lip reading in the wild*, in Asian Conf. Comput. Vis., pp. 87–103, Springer, (2016).
- [7] J. S. Chung and A. Zisserman, *Lip Reading Sentences in the Wild*, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 6447–6456, (2017). doi: 10.48550/arXiv.1611.05358.
- [8] Y. Lu and H. Li, *Automatic Lip-Reading System Based on Deep Convolutional Neural Network and Attention-Based Long Short-Term Memory*, Applied Sciences, vol. 9, no. 8, Art. no. 1599, (2019). doi: 10.3390/app9081599.
- [9] A. Mesbah, A. Berrahou, H. Hammouchi, H. Berbia, H. Qjidaa, and M. Daoudi, *Lip reading with Hahn Convolutional Neural Networks*, Image and Vision Computing, vol. 88, pp. 76–83, (2019). doi: 10.1016/j.imavis.2019.05.003.
- [10] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, *LipNet: End-to-End Sentence-Level Lipreading*, arXiv preprint arXiv:1611.01599, (2016). doi: 10.48550/arXiv.1611.01599.
- [11] H. Prifti and Bekir Karlik, *Recognition of GUI Web Elements using the Deep Learning Approach and Yolo Architecture*, Systems and Computing, vol. 2, no. 1, Jan. (2026), doi: 10.64409/sycom.v2.i1.28.

- [12] S. Ji, W. Xu, M. Yang, and K. Yu, *3D convolutional neural networks for human action recognition*, IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 1, pp. 221–231, (2013). doi: 10.1109/TPAMI.2012.59.
- [13] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, *Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks*, in Proc. Int. Conf. Mach. Learn. (ICML), (2006). doi: 10.1145/1143844.1143891.
- [14] V. Kazemi and J. Sullivan, *One millisecond face alignment with an ensemble of regression trees*, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 1867–1874, (2014). doi: 10.1109/CVPR.2014.241.
- [15] D. Li, Y. Gao, C. Zhu, Q. Wang, and R. Wang, *Improving Speech Recognition Performance in Noisy Environments by Enhancing Lip Reading Accuracy*, Sensors, vol. 23, no. 4, Art. no. 2053, (2023). doi: 10.3390/s23042053.
- [16] Y. Liu, C. Lin, M. Wang, S. Liang, Z. Chen, and L. Chen, *End-to-End Chinese Lip-Reading Recognition Based on Multi-modal Fusion*, in Proc. ICFTIC, pp. 794–801, 2022.
- [17] X. Yu, L. Wang, C. Chen, J. Tie, and S. Guo, *Multimodal Learning of Audio-Visual Speech Recognition with Liquid State Machine*, in Neural Information Processing, pp. 552–563, Springer, 2023.
- [18] W. Qiang et al., *Advancing Lip-Reading for Unseen Speakers Through Fusion and Augmentation of Spatio-temporal Landmarks and Visual Features*, in Neural Information Processing, pp. 268–282, Springer, 2025.
- [19] S. Jeon, A. Elsharkawy, and M. S. Kim, *Lipreading Architecture Based on Multiple Convolutional Neural Networks for Sentence-Level Visual Speech Recognition*, Sensors, vol. 22, no. 1, p. 72, (2022). doi: 10.3390/s22010072.
- [20] U. Tabassum, A. S. Barakey, L. P. Rao, and M. C. Felcita, *SILENT2SOUND: Visual-to-Sound Audio Speech Reconstruction Using Lip-Reading AI*, IRJMETS, 2023.
- [21] M. Cooke, J. Barker, S. Cunningham, and X. Shao, *An audio-visual corpus for speech perception and automatic speech recognition*, J. Acoust. Soc. Am., vol. 120, no. 5, pp. 2421–2424, (2006). doi: 10.1121/1.2229005.
- [22] F. Xue, T. Yang, K. Liu, Z. Hong, M. Cao, D. Guo, and R. Hong, *LCSNet: End-to-end Lipreading with Channel-aware Feature Selection*, ACM Transactions on Multimedia Computing, Communications and Applications, vol. 19, no. 1s, pp. 1–21, (2023).
- [23] H. Huang, C. Song, J. Ting, T. Tian, C. Hong, Z. Di, and D. Gao, *A Novel Machine Lip Reading Model*, Procedia Computer Science, (2022). doi: 10.1145/3571722.
- [24] Y. Shuang and Z. Yuanhang and F. Dalu and Y. Mingmin and W. Chenhao and X. Jingyun and L. Keyu and S. Shiguang and C. Xilin, *LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild*, 14th IEEE international conference on automatic face & Gesture Recognition (FG 2019), pp. 1–8, (2019). doi: 10.1109/FG.2019.8756572.
- [25] T. Afouras, J. Son Chung, A. Senior, O. Vinyals, and A. Zisserman. *Deep Audio-Visual Speech Recognition*. ArXiv e-prints, September, (2018). doi: 10.48550/arXiv.1809.02108.

Author information

H. Benaliouche, Lire Laboratory, Faculty of New Technologies of Information and Communications (NTIC), Department of Fundamental Computer science and its applications, University Abdelhamid Mehri Constantine 2, Algeria.

E-mail: houda.benaliouche@univ-constantine2.dz

D. Bellibel, Faculty of New Technologies of Information and Communications (NTIC), Department of Fundamental Computer science and its applications, University Abdelhamid Mehri Constantine 2, Algeria.

E-mail: douaa.bellibel@univ-constantine2.dz

M. Habbati, Faculty of New Technologies of Information and Communications (NTIC), Department of Fundamental Computer science and its applications, University Abdelhamid Mehri Constantine 2, Algeria.

E-mail: malak.habbati@univ-constantine2.dz