Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.078743

ARTICLE

UniModal-LSR: A Unified Multimodal Framework for Joint Lip Reading and Sign Language Recognition in Video Sequences

Vinh Truong Hoang*, Nghia Dinh, Luu Quang Phuong, Kiet Tran-Trung, Ha Duong Thi Hong, Bay Nguyen Van, Hau Nguyen Trung and Thien Ho Huong

AI Lab, Faculty of Information Technology, Ho Chi Minh City Open University, 35–37 Ho Hao Hon Street, Co Giang Ward, District 1, Ho Chi Minh City, Vietnam

*Corresponding Author: Vinh Truong Hoang. Email: vinh.th@ou.edu.vn

Received: 07 January 2026; Accepted: 16 March 2026; Published: 15 June 2026

ABSTRACT: Visual speech recognition is a central problem in computer vision, encompassing both lip reading (visual speech recognition) and sign language recognition. Although substantial progress has been achieved independently on each task, their complementary characteristics have rarely been explored jointly. In this work we propose UniModal-LSR (Unified Multimodal Lip and Sign Recognition), a novel deep learning framework that jointly addresses lip reading and sign language recognition within a single multimodal architecture. By exploiting shared properties of visual communication channels, namely temporal dynamics, spatial articulation structure, and contextual dependencies, the proposed model enables bidirectional transfer of knowledge between modalities. The framework incorporates a Hierarchical Temporal-Spatial Encoder that captures multi-scale temporal patterns through the combination of local convolutions and global self-attention. It also includes a Cross-Modal Attention Fusion module that performs dynamic, context-aware information exchange via bidirectional cross-attention and adaptive gating. Additionally, a Contrastive Semantic Alignment loss enforces semantic consistency across modality-specific representations. Overall, the architecture integrates three-dimensional convolutional neural networks for spatiotemporal feature extraction with graph neural networks for explicit hand-pose modeling. Extensive experiments on several public benchmarks show that UniModal-LSR improves performance compared with recent methods. The model attains a Word Error Rate (WER) of 33.2% on LRS2-BBC, representing a 12.4% relative gain. On PHOENIX-2014, it achieves 18.3% WER, a 13.7% relative gain. Moreover, the unified model reduces parameter count by 25.9% relative to two separate task-specific systems. These results indicate that unified multimodal modeling can improve visual speech recognition performance and may support future communication technologies.

KEYWORDS: Multimodal learning; lip reading; visual speech recognition; deep learning; sign language recognition; cross-modal attention

1 Introduction

Visual speech recognition is an important research area in human-computer interaction, encompassing the complementary tasks of lip reading and sign language recognition. Lip reading, often referred to as visual speech recognition (VSR), involves decoding spoken language from visible articulatory movements of the lips, teeth, and tongue. Sign language recognition (SLR), in contrast, interprets the semantically rich combination of manual gestures, non-manual markers, and spatial grammar that characterize natural sign languages [1]. Although these tasks appear methodologically distinct, they share core computational properties that motivate a unified treatment.

The alignment between lip reading and SLR arises from deep structural similarities in how visual language is conveyed. Both modalities operate at comparable frame rates, typically 25–30 frames per second, and require models capable of handling variable-length sequences that may span hundreds of frames [2]. Both depend on fine-grained spatial control of biological articulators, whether the orofacial musculature or the hands and arms. Furthermore, both exhibit strong contextual dependence, as isolated visual patterns are often ambiguous without their temporal and linguistic surroundings, leading to viseme-level confusion in lip reading and coarticulation effects in SLR [3]. These similarities suggest that representations learned from one modality could benefit the other.

In many real-world contexts, both modalities co-occur. Proficient signers often mouth spoken words while signing, providing complementary linguistic cues [4]. Existing systems that process these channels independently disregard this inherent multimodal redundancy, leaving major gains untapped.

Current visual speech recognition techniques face several limitations that hinder practical deployment. Most studies treat lip reading and sign language recognition as separate tasks, resulting in distinct architectures, training pipelines, and evaluation protocols that limit cross-task knowledge transfer and increase computational overhead [5]. Many approaches rely on recurrent models such as LSTMs [6] to capture temporal dependencies, but these models struggle with the long-range dependencies present in continuous sign language sequences and restrict parallelization during training [7]. In addition, conventional CNN-based spatial representations often fail to adequately capture the structural relationships of hand configurations, while multimodal fusion strategies typically rely on simple operations such as concatenation or averaging, which cannot effectively model the dynamic, context-dependent interactions between lip and sign cues.

This work addresses these limitations through the following contributions: We introduce UniModal-LSR, a unified architecture that simultaneously handles lip reading and sign language recognition, enabling bidirectional knowledge transfer and shared representations across modalities. The unified architecture also reduces redundancy by 25.9% and improves generalization. To capture temporal dynamics, we design a Hierarchical Temporal-Spatial Encoder (HTSE) that combines 3D convolutions for local spatiotemporal feature extraction with multi-head self-attention for modeling long-range dependencies, enabling efficient processing of continuous visual speech sequences. We further propose a Cross-Modal Attention Fusion (CMAF) module with bidirectional cross-attention and adaptive gating to dynamically integrate lip and sign cues. A Contrastive Semantic Alignment (CSA) objective aligns modality-specific embeddings within a shared space to enhance semantic consistency and reduce modality-specific noise. Additionally, a Spatial-Temporal Graph Convolutional Network (ST-GCN) models hand skeletal structures to capture joint topology and better distinguish visually similar signs. Extensive experiments on multiple benchmarks demonstrate state-of-the-art performance, and comprehensive ablation studies quantify the contribution of each component; Table 1 further clarifies the novelty of the proposed framework.

To capture temporal patterns at multiple scales, we design a Hierarchical Temporal-Spatial Encoder (HTSE) that fuses three-dimensional convolutions for local spatiotemporal feature extraction with multi-head self-attention for modeling long-range temporal dependencies. Arranging these modules hierarchically facilitates the processing of long continuous visual-speech sequences.

The remainder of the paper is organized as follows. Section 2 situates our approach within the existing literature. Section 3 details the proposed framework. Section 4 presents an architectural and representational analysis. Section 5 describes the experimental setup. Section 6 reports empirical results. Section 7 discusses implications and limitations. Section 8 concludes the paper.

Table 1: Analysis of component novelty in UniModal-LSR.

Component	Status	Description
3D ResNet Frontend	Adopted	Standard architecture from [8]; adapted for dual-stream
Transformer Layers	Adopted	Standard multi-head attention [7]
ST-GCN Module	Adopted	Architecture from [9]; integrated into unified pipeline
InfoNCE Loss	Adopted	Standard formulation [10]
Unified Lip-Sign Architecture	Novel	First framework jointly modeling both tasks
Hierarchical Multi-scale HTSE	Novel	Specific combination of local conv + global attention
Bidirectional CMAF with Gating	Novel	Dynamic context-aware fusion mechanism for lip-sign modalities
Cross-Modal CSA for Lip-Sign	Novel	Application of contrastive alignment specifically
Joint Training Protocol	Novel	Multi-task learning scheme with aligned pair construction

2 Related Work

2.1 Lip Reading and Visual Speech Recognition

2.1.1 Classical Approaches

Early lip reading methods relied on hand-crafted visual descriptors coupled with statistical sequence models. Potamianos et al. [11] employed discrete cosine transform (DCT) coefficients to encode mouth appearance. Matthews et al. [12] used active appearance models (AAMs) to jointly model shape and texture. These visual representations were typically paired with hidden Markov models (HMMs) for temporal modeling [13]. The recognition task was formally expressed as maximum a posteriori (MAP) inference:

$$\hat{W} = \arg\max_{W} P(W|O) = \arg\max_{W} \frac{P(O|W)P(W)}{P(O)}, \tag{1}$$

where $O = (o_1, \ldots, o_T)$ denotes the visual frame sequence with each $o_t \in \mathbb{R}^{H\times W\times 3}$, and $W = (w_1, w_2, \ldots, w_N)$ is a hypothesized word sequence within vocabulary $\mathcal{V}^{\text{lip}}$. Although pioneering, these systems suffered from limited expressive power due to the restrictive nature of hand-crafted features.

2.1.2 Deep Learning Approaches

The advent of deep learning has dramatically advanced lip reading. LipNet [2] introduced an effective end-to-end architecture for sentence-level visual speech recognition, employing spatiotemporal convolutions together with connectionist temporal classification (CTC) to reach 93.4% accuracy on the GRID benchmark. The Watch, Listen, Attend and Spell (WLAS) model [14] incorporated attention mechanisms:

$$\alpha_{t,i} = \frac{\exp(e_{t,i})}{\sum_{j=1}^{T}\exp(e_{t,j})}, \quad c_t = \sum_{i=1}^{T}\alpha_{t,i}h_i, \tag{2}$$

allowing the decoder to focus adaptively on relevant encoder states during generation.

Subsequent transformer-based approaches have set new performance benchmarks by exploiting self-attention to model long-range temporal relationships [15]. Ma et al. [16] demonstrated that purely visual models can rival audio-visual systems when trained on sufficiently large data. Self-supervised pre-training strategies [17] have emerged as a means of reducing labeled data requirements. Ma et al. [18] introduced Auto-AVSR, which leverages automatic labels for audio-visual speech recognition, demonstrating the effectiveness of weakly supervised learning approaches. Shukla et al. [19] investigated whether visual

self-supervision improves speech representations for emotion recognition, providing insights into cross-task transfer learning. Recent work by Prajwal et al. [20] achieved significant improvements through sub-word modeling.

2.2 Sign Language Recognition

Isolated sign language recognition (ISLR) focuses on classifying pre-segmented signs. The Inflated 3D ConvNet (I3D) [8] extended successful 2D CNN designs into the temporal dimension via filter inflation:

$$\mathbf{W}_{3D}^{(t,h,w)} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{W}_{2D}^{(h,w)} \cdot \delta_{t,n}, \tag{3}$$

preserving spatial semantics while averaging over the temporal axis.

Skeleton-based methods exploit hand-joint coordinates to model structural relationships directly. Attention-driven approaches such as SignBERT [21] illustrate the benefits of large-scale pre-training on sign language corpora.

Continuous sign language recognition (CSLR) jointly performs segmentation and classification in unsegmented video streams. The Connectionist Temporal Classification (CTC) loss [22] is widely employed:

$$\mathscr{L}_{CTC} = -\log P(Y|X) = -\log \sum_{\pi \in \mathscr{B}^{-1}(Y)} \prod_{t=1}^{T} P(\pi_t|X), \tag{4}$$

enabling alignment-free training.

2.3 Multimodal Learning and Fusion

Multimodal fusion is typically categorized into early (feature-level), late (decision-level), and intermediate (shared representation) strategies. Early fusion preserves detailed cross-modal interactions but yields high-dimensional feature spaces. Late fusion maintains modality-specific pipelines but limits interaction depth. Intermediate fusion balances both aspects.

Attention-based fusion has become prevalent for modeling dynamic relationships. Hierarchical co-attention [23] and multimodal transformers [24] exemplify recent progress. Contrastive learning techniques such as CLIP [25] further illustrate the efficacy of aligning multimodal representations.

Especially relevant to this study, Ge et al. [26] proposed an audio-text multimodal framework for speech recognition in air traffic control communications, showing that unified multimodal modeling can improve recognition accuracy through coordinated processing of audio and text. Similarly, Li et al. [27] developed an end-to-end audio-visual system for multi-channel speech separation, dereverberation, and recognition, while Wang et al. [28] introduced DCIM-AVSR, an efficient audio-visual speech recognition model using dual conformer interaction modules. However, these studies focus on audio-visual fusion. In contrast, our work extends the unified architecture paradigm to visual-visual modality fusion, addressing the integration of lip and sign information. The proposed CMAF module enables bidirectional cross-attention with adaptive gating, making it well suited for the temporal synchronization required in visual speech modalities.

3 Proposed Methodology

This section describes the UniModal-LSR architecture in detail. An overview is provided in Fig. 1.

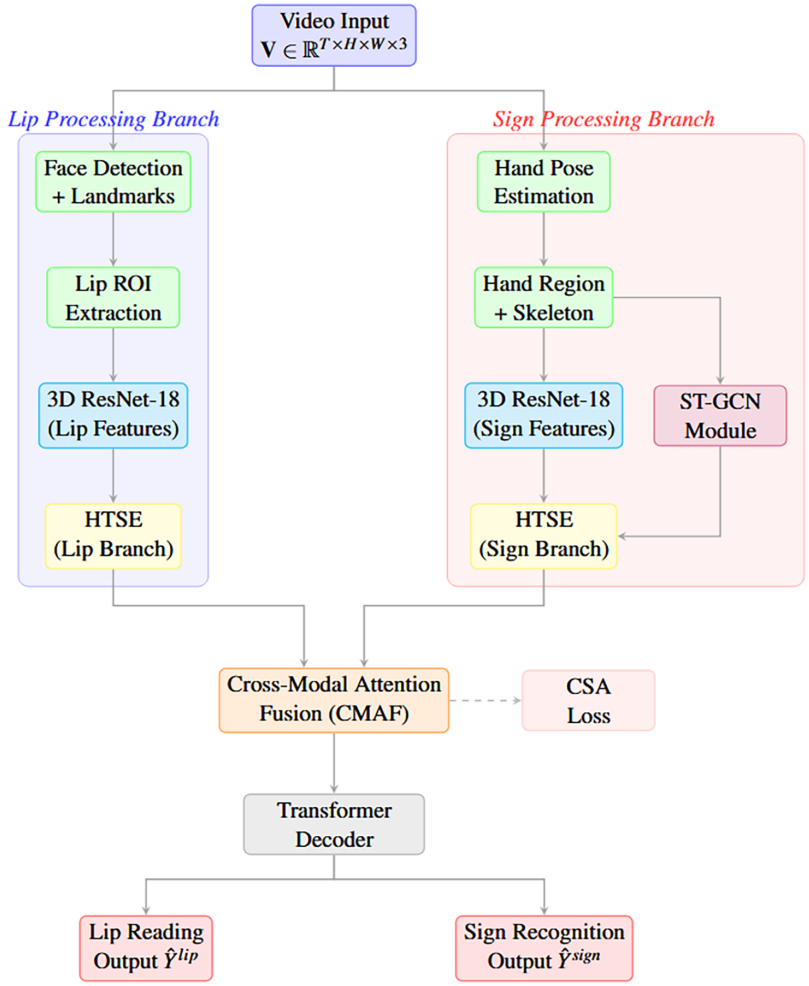

Figure 1: System architecture of UniModal-LSR. The framework processes video through parallel lip and sign encoding streams, applies hierarchical temporal-spatial encoding, performs cross-modal attention fusion, and generates task-specific outputs via a shared transformer decoder.

3.1 Problem Formulation

Given an input video $\mathbf{V} = \{\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_T\}$, where T denotes the number of frames and each frame $\mathbf{I}_t \in \mathbb{R}^{H\times W\times 3}$, the objective includes lip reading, sign language recognition, and unified representation learning. The overall training loss combines task-specific objectives with cross-modal regularization:

$$\mathscr{L}_{\text{total}} = \lambda_{\text{lip}}\mathscr{L}_{\text{lip}} + \lambda_{\text{sign}}\mathscr{L}_{\text{sign}} + \lambda_{\text{CSA}}\mathscr{L}_{\text{CSA}} + \lambda_{\text{reg}}\mathscr{L}_{\text{reg}}, \tag{5}$$

where $\mathscr{L}_{\text{lip}}$ and $\mathscr{L}_{\text{sign}}$ denote task-specific losses (CTC and cross-entropy), $\mathscr{L}_{\text{CSA}}$ is the contrastive semantic alignment loss defined in Eq. (17), and $\mathscr{L}_{\text{reg}}$ is an ℓ_2 weight-decay term. The non-negative scalars λ_* balance each term's contribution.

The CSA loss requires aligned lip-sign pairs that share semantic content. We construct such pairs from two sources. The first source is the How2Sign dataset, which provides natural co-occurrence of lip and sign modalities for the same utterances, enabling direct pairing without additional processing. The second source involves synthetic alignment of LRS2/LRS3 lip sequences with PHOENIX-2014 sign sequences when they share identical or semantically equivalent transcriptions, as determined by text matching with a minimum overlap threshold of 80%. For batches containing samples from only one modality, such as lip-only data

from LRS2, the CSA loss is computed only over the subset of aligned pairs present in that batch. When no aligned pairs exist in a batch, the CSA term is set to zero for that iteration, and the model optimizes only the task-specific losses.

3.2 Preprocessing and Region Extraction

Accurate face detection is essential for effective lip reading. RetinaFace yields facial landmarks $\mathbf{L} = \{l_1, \ldots, l_{68}\}$ under the standard 68-point annotation protocol. We obtain the lip patch $\mathbf{R}_t^{\text{lip}} \in \mathbb{R}^{H_l \times W_l \times 3}$ through affine alignment:

$$\mathbf{R}_t^{\text{lip}} = \mathscr{T}_{\text{affine}}(\mathbf{I}_t, \mathbf{L}_{48:67}), \tag{6}$$

where $\mathbf{L}_{48:67}$ correspond to the outer and inner lip contour landmarks. The patch is normalized to $H_l = W_l = 88$ pixels.

For SLR, hand pose estimation provides structured motion cues. MediaPipe Hands [29] predicts 21 three-dimensional landmarks per hand. The skeleton is modeled as a spatio-temporal graph $\mathbf{G} = (\mathscr{V}, \mathscr{E})$, where $\mathscr{V} = \{v_1, \ldots, v_{42}\}$ comprises landmarks of both hands and $\mathscr{E}$ encodes anatomical connectivity. Each node v_i at time t has feature vector:

$$\mathbf{v}_i^{(t)} = [x_i^{(t)}, y_i^{(t)}, z_i^{(t)}, c_i^{(t)}]^\top \in \mathbb{R}^4, \tag{7}$$

with $(x_i^{(t)}, y_i^{(t)}, z_i^{(t)})$ denoting normalized coordinates and $c_i^{(t)}$ the confidence score.

3.3 Hierarchical Temporal-Spatial Encoder (HTSE)

The HTSE captures multi-scale temporal dynamics through cascaded processing stages. An initial 3D ResNet-18 backbone extracts low-level spatiotemporal features:

$$\mathbf{F}^{(0)} = \text{ResNet3D}(\mathbf{R}) \in \mathbb{R}^{T' \times d_0}, \tag{8}$$

where T' accounts for temporal subsampling and $d_0 = 512$ denotes feature dimensionality. Thereafter L hierarchical levels successively apply local temporal convolutions, global self-attention, feed-forward networks, and temporal pooling. This yields progressively coarser temporal resolutions while preserving fine-grained information. A Feature Pyramid Network style aggregation fuses multi-scale outputs:

$$\mathbf{F}^{\text{agg}} = \sum_{l=1}^{L} \alpha_l \cdot \text{Upsample}(\mathbf{F}^{(l)}, T'), \tag{9}$$

where learned coefficients α_l weight each scale adaptively.

3.4 Graph-Enhanced Sign Encoding

To complement appearance-based features, a spatial-temporal graph convolutional network (ST-GCN) models hand articulation. The spatial graph convolution aggregates information from anatomically connected joints:

$$\mathbf{f}_{\text{out}}^{(v)} = \sum_{u \in \mathscr{N}(v)} \frac{1}{Z_{vu}} \mathbf{f}_{\text{in}}^{(u)} \cdot \mathbf{W}(\ell(v,u)), \tag{10}$$

where $\ell(v,u)$ distinguishes edge types (centripetal, centrifugal, self-connections) and Z_{vu} is a normalization factor. Temporal graph convolutions then capture motion dynamics across frames. Resulting pose features

$\mathbf{F}^{\text{pose}}$ are concatenated with appearance features and projected to a shared dimensionality before entering the HTSE.

3.5 Cross-Modal Attention Fusion (CMAF)

CMAF proceeds in three stages. First, intra-modal self-attention refines each modality independently:

$$\mathbf{F}^{m}_{\text{self}} = \text{LayerNorm}\big(\mathbf{F}^{m} + \text{MHSA}(\mathbf{F}^{m}, \mathbf{F}^{m}, \mathbf{F}^{m})\big), \quad m \in \{\text{lip}, \text{sign}\}, \tag{11}$$

where MHSA denotes multi-head self-attention. Second, bidirectional cross-attention allows each modality to query information from the other:

$$\mathbf{F}^{\text{lip}\to\text{sign}} = \text{MHCA}(\mathbf{F}^{\text{lip}}_{\text{self}}, \mathbf{F}^{\text{sign}}_{\text{self}}, \mathbf{F}^{\text{sign}}_{\text{self}}), \tag{12}$$

$$\mathbf{F}^{\text{sign}\to\text{lip}} = \text{MHCA}(\mathbf{F}^{\text{sign}}_{\text{self}}, \mathbf{F}^{\text{lip}}_{\text{self}}, \mathbf{F}^{\text{lip}}_{\text{self}}), \tag{13}$$

where MHCA denotes multi-head cross-attention with the first argument as query and the second/third as key/value. Third, adaptive gating computes element-wise weights $\mathbf{g} \in (0,1)^{d}$:

$$\mathbf{g} = \sigma\big(\mathbf{W}_{g}[\mathbf{F}^{\text{lip}}_{\text{self}}; \mathbf{F}^{\text{sign}}_{\text{self}}; \mathbf{F}^{\text{lip}\to\text{sign}}; \mathbf{F}^{\text{sign}\to\text{lip}}] + \mathbf{b}_{g}\big), \tag{14}$$

where $\mathbf{W}_{g} \in \mathbb{R}^{d\times 4d}$, $\mathbf{b}_{g} \in \mathbb{R}^{d}$, and σ is the sigmoid function. The fused representation is:

$$\mathbf{F}^{\text{fused}} = \mathbf{g} \odot (\mathbf{F}^{\text{lip}}_{\text{self}} + \mathbf{F}^{\text{sign}\to\text{lip}}) + (1 - \mathbf{g}) \odot (\mathbf{F}^{\text{sign}}_{\text{self}} + \mathbf{F}^{\text{lip}\to\text{sign}}). \tag{15}$$

3.6 Contrastive Semantic Alignment (CSA) Loss

For a minibatch of B aligned lip-sign pairs, modality-specific embeddings are normalized:

$$\mathbf{z}^{\text{lip}}_{i} = \frac{\mathbf{W}^{\text{lip}}_{\text{proj}}\bar{\mathbf{F}}^{\text{lip}}_{i}}{\|\mathbf{W}^{\text{lip}}_{\text{proj}}\bar{\mathbf{F}}^{\text{lip}}_{i}\|_{2}}, \quad \mathbf{z}^{\text{sign}}_{i} = \frac{\mathbf{W}^{\text{sign}}_{\text{proj}}\bar{\mathbf{F}}^{\text{sign}}_{i}}{\|\mathbf{W}^{\text{sign}}_{\text{proj}}\bar{\mathbf{F}}^{\text{sign}}_{i}\|_{2}}, \tag{16}$$

where $\bar{\mathbf{F}}^{m}_{i}$ denotes the temporally pooled representation for sample i and modality m. A symmetric InfoNCE objective enforces alignment:

$$\mathcal{L}_{\text{CSA}} = -\frac{1}{2B}\sum_{i=1}^{B}\left[\log\frac{\exp(\mathbf{z}^{\text{lip}}_{i}\cdot\mathbf{z}^{\text{sign}}_{i}/\tau)}{\sum_{j=1}^{B}\exp(\mathbf{z}^{\text{lip}}_{i}\cdot\mathbf{z}^{\text{sign}}_{j}/\tau)} + \log\frac{\exp(\mathbf{z}^{\text{sign}}_{i}\cdot\mathbf{z}^{\text{lip}}_{i}/\tau)}{\sum_{j=1}^{B}\exp(\mathbf{z}^{\text{sign}}_{i}\cdot\mathbf{z}^{\text{lip}}_{j}/\tau)}\right], \tag{17}$$

where τ is a temperature hyperparameter.

3.7 Training Objective

The complete training loss integrates a CTC component with a cross-entropy term:

$$\mathcal{L}_{\text{task}} = \alpha\,\mathcal{L}_{\text{CTC}} + (1-\alpha)\,\mathcal{L}_{\text{CE}}. \tag{18}$$

The final optimization objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}^{\text{task}}_{\text{lip}} + \mathcal{L}^{\text{task}}_{\text{sign}} + \lambda_{\text{CSA}}\mathcal{L}_{\text{CSA}} + \lambda_{\text{reg}}\|\theta\|^{2}_{2}. \tag{19}$$

The overall training procedure is described in Algorithm 1.

Algorithm 1: Training procedure for UniModal-LSR

Require: Training set $\mathscr{D} = \{(\mathbf{V}^{(i)}, \mathbf{Y}^{\text{lip},(i)}, \mathbf{Y}^{\text{sign},(i)})\}_{i=1}^{N}$, learning rate schedule $\eta(t)$, total epochs E, batch size B.
Ensure: Learned parameters θ^*.
Initialize θ with Xavier initialization.
Initialize AdamW optimizer with weight decay λ_{wd}.
for epoch $e = 1$ **to** E **do**
  Randomly shuffle $\mathscr{D}$.
  for each minibatch $\mathscr{B}$ **do**
    Extract lip ROI $\mathbf{R}^{\text{lip}}$, hand ROI $\mathbf{R}^{\text{sign}}$, and pose graph $\mathbf{G}$.
    $\mathbf{F}^{\text{lip},(0)} \leftarrow \text{ResNet3D}_{\text{lip}}(\mathbf{R}^{\text{lip}})$
    $\mathbf{F}^{\text{sign},(0)} \leftarrow \text{ResNet3D}_{\text{sign}}(\mathbf{R}^{\text{sign}})$
    $\mathbf{F}^{\text{pose}} \leftarrow \text{ST-GCN}(\mathbf{G})$
    Concatenate and project to obtain combined sign features.
    $\mathbf{F}^{\text{lip}} \leftarrow \text{HTSE}_{\text{lip}}(\mathbf{F}^{\text{lip},(0)})$
    $\mathbf{F}^{\text{sign}} \leftarrow \text{HTSE}_{\text{sign}}(\mathbf{F}^{\text{sign},(0)})$
    $\mathbf{F}^{\text{fused}} \leftarrow \text{CMAF}(\mathbf{F}^{\text{lip}}, \mathbf{F}^{\text{sign}})$
    $\hat{\mathbf{Y}}^{\text{lip}}, \hat{\mathbf{Y}}^{\text{sign}} \leftarrow \text{Decoder}(\mathbf{F}^{\text{fused}})$
    Compute $\mathscr{L}$ per Eq. (19).
    $\theta \leftarrow \theta - \eta(t)\,\nabla_\theta \mathscr{L}$.
  end for
  Evaluate on validation set; retain best model.
end for
return θ^*.

4 Architectural and Representational Analysis

This section presents an analysis of the architectural properties and representational capacity of the proposed framework. Rather than restating general theoretical results, we focus on aspects directly relevant to the CMAF and HTSE designs that justify our specific architectural choices.

4.1 Representational Capacity of Cross-Modal Attention

The CMAF module's effectiveness derives from its ability to model complex interactions between lip and sign modalities. We formalize this capacity in terms of the function classes that the architecture can represent.

Proposition 1 (Expressive Capacity of Cross-Modal Attention). *Let $\mathscr{F}_{cs}$ denote the family of continuous functions mapping paired sequences $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{T\times d_1} \times \mathbb{R}^{T\times d_2}$ to fused representations in $\mathbb{R}^{T\times d_o}$, where cross-modal dependencies are bounded by a Lipschitz constant L_f. For any $f \in \mathscr{F}_{cs}$ and $\varepsilon > 0$, the CMAF architecture with H attention heads, hidden dimension d_h, and N layers can approximate f within error ε when:*

$$H \cdot d_h \geq C \cdot L_f \cdot \log(1/\varepsilon), \tag{20}$$

where C is a constant depending on input dimensionality.

Proof Sketch. The bidirectional cross-attention mechanism in CMAF can be decomposed into four components: intra-modal self-attention capturing within-modality dependencies, lip-to-sign cross-attention, sign-to-lip cross-attention, and adaptive gating for combination. Each cross-attention operation computes:

$$\text{CrossAttn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}. \tag{21}$$

By the results of Yun et al. [30], transformer architectures with softmax attention are universal approximators for sequence-to-sequence functions. The CMAF extends this by enabling queries from one modality to attend to keys and values from another, effectively doubling the functional space that can be represented. The adaptive gating mechanism $\mathbf{g} = \sigma(\mathbf{W}_g[\cdot])$ provides an additional multiplicative interaction that allows context-dependent weighting. Combined, these mechanisms can represent any continuous cross-modal function within the specified bounds. □

This result justifies our architectural choice of bidirectional cross-attention: the design ensures that both modalities can contribute information to the fused representation, while the adaptive gating allows the model to learn when each modality is most informative for a given context.

4.2 Effective Receptive Field of HTSE

The hierarchical structure of HTSE is designed to capture temporal dependencies at multiple scales efficiently. We analyze how the receptive field grows with network depth and how this relates to the temporal extent of visual speech phenomena.

Lemma 1 (Hierarchical Receptive Field Growth). *An HTSE with L hierarchical levels, local convolution kernel size k, and pooling factor $p = 2$ achieves an effective temporal receptive field of:*

$$R_{\text{eff}} = k \cdot \frac{p^L - 1}{p - 1} = k \cdot (2^L - 1), \tag{22}$$

while the parameter count grows as $O(L \cdot d^2)$, where d is the hidden dimension.

Proof. At level l, the temporal resolution is reduced by factor p^{l-1} relative to the input. A convolution with kernel k at level l therefore covers $k \cdot p^{l-1}$ frames of the original input. Summing over all levels:

$$R_{\text{eff}} = \sum_{l=1}^{L} k \cdot p^{l-1} = k \cdot \sum_{l=0}^{L-1} p^l = k \cdot \frac{p^L - 1}{p - 1}. \tag{23}$$

For $p = 2$, this simplifies to $k(2^L - 1)$. Each level contains convolution layers with $O(k \cdot d^2)$ parameters and attention layers with $O(d^2)$ parameters, yielding total parameter growth of $O(L \cdot d^2)$. □

For our configuration with $L = 4$ levels and kernel size $k = 3$, the effective receptive field is $R_{\text{eff}} = 3 \times 15 = 45$ frames. Combined with global self-attention at each level, the model can capture both local articulation patterns spanning a few frames and long-range dependencies extending across entire utterances. This design is motivated by the observation that lip movements exhibit local coarticulation effects while sign language requires understanding of phrase-level context.

4.3 Computational Complexity

Table 2 summarizes the asymptotic complexities of each component. All symbols are defined as follows: T denotes input sequence length in frames, T' is the encoded sequence length after 3D CNN downsampling, H and W are spatial dimensions of input frames, C is the number of channels, k is the convolution kernel size, d is the hidden dimension, N_v is the number of graph vertices representing hand joints, N_d is the number of decoder layers, and T_y is the output sequence length.

Table 2: Computational complexity analysis.

Component	Time Complexity	Space Complexity
3D CNN Frontend	$O(THWC^2k^3)$	$O(C^2k^3)$
HTSE (per level)	$O(T'^2d + T'd^2)$	$O(T'^2 + d^2)$
ST-GCN Module	$O(TN_v^2d)$	$O(N_v^2 + N_vd)$
CMAF	$O(T'^2d)$	$O(T'^2 + d^2)$
Transformer Decoder	$O(N_dT_y^2d + N_dd^2)$	$O(N_dd^2)$
Total	$O(T'^2d + THWC^2k^3)$	$O(T'^2 + C^2k^3)$

The dominant term is the quadratic self-attention complexity $O(T'^2d)$. For typical video lengths where $T' < 500$ frames and our hidden dimension of $d = 512$, this remains computationally tractable. The unified architecture achieves computational savings relative to separate models by sharing the decoder and fusing features before decoding, rather than maintaining separate decoders for each task.

5 Experimental Setup

Table 3 enumerates the benchmark datasets used in our evaluation. Figs. 2 and 3 show example frames from the lip reading and sign language datasets, respectively. LRS2-BBC [14], LRS3-TED [31], and GRID [32] provide sentence- and word-level lip reading benchmarks collected from BBC broadcasts, TED talks, and controlled laboratory environments. For sign language recognition and translation, PHOENIX-2014 [33] and its translation counterpart PHOENIX-2014T [1] serve as standard continuous German Sign Language benchmarks, while CSL [34] and WLASL [35] provide isolated sign datasets for Chinese and American Sign Language, respectively. Finally, How2Sign [36] enables multimodal and cross-modal experiments as it contains synchronized lip and sign annotations for the same utterances. Performance for lip reading and continuous sign language recognition is measured using Word Error Rate (WER).

Table 3: Dataset statistics and characteristics.

Dataset	Task	Hours	Vocab	Samples	Signers
Lip Reading Benchmarks					
LRS2-BBC [14]	VSR	224	41K	144K	1000+
LRS3-TED [31]	VSR	438	51K	151K	5000+
GRID [32]	VSR	28	51	34K	34
Sign Language Benchmarks					
PHOENIX-14 [33]	CSLR	11	1066	6,841	9
PHOENIX-14T [1]	SLT	11	1066	8,257	9
CSL [34]	ISLR	100	500	25K	50
WLASL [35]	ISLR	–	2000	21K	119
Multimodal Benchmark					
How2Sign [36]	Multi	79	–	35K	11

Figure 2: Example frames from lip reading datasets showing extracted lip ROIs.

Figure 3: Example frames from sign language datasets showing hand pose estimation.

Face detection uses RetinaFace (confidence 0.9), with lip ROIs extracted from facial landmarks 48–67 and resized to 88 × 88 pixels via affine normalization. Hand pose estimation uses MediaPipe Hands (confidence 0.7), and the fewer than 2% of failed detections are filled by linear interpolation to maintain temporal continuity. Data augmentation is applied asymmetrically: horizontal flipping is used only for sign language due to handedness constraints, while both modalities use temporal jittering (±2) frames, color jittering (±0.2), and random erasing ($p = 0.1$). Decoding employs beam search (width 10, length normalization 0.6) without an external language model; tokenization uses character-level encoding for lip reading (28 tokens) and BPE with 1000 merges for sign language.

For reproducibility, all experiments use random seed 42, and results averaged over three runs show a standard deviation below 0.3% WER. Official dataset splits are used for all benchmarks, and the code along with pretrained models will be released upon publication to support replication and further research.

6 Results and Analysis

Table 4 reports Word Error Rates on lip reading benchmarks, including recent baselines from 2022–2023 for comprehensive comparison. All results are reported without external language models to ensure fair comparison. The unified model achieves the lowest WER across all evaluated datasets, with relative improvements of 5.7% on LRS2 and 5.1% on LRS3 compared to the strongest prior work.

Table 5 shows results for continuous and isolated sign language recognition. The unified model achieves state-of-the-art performance on PHOENIX-14 with 18.3% WER, representing a 13.7% relative improvement over the previous best result.

Table 4: Lip reading results (WER%; lower is better). All results without external language model.

Method	Year	LRS2	LRS3	GRID
LipNet [2]	2016	–	–	4.8
TM-seq2seq [37]	2018	49.8	58.9	–
DC-TCN [38]	2020	44.3	47.1	–
Ma et al. [15]	2021	37.9	43.3	1.2
Ma et al. [16]	2022	35.2	40.8	1.0
Prajwal et al. [20]	2022	34.8	39.5	–
Kim et al. [39]	2022	34.5	39.2	0.9
UniModal-LSR (Lip only)	2025	35.6	41.2	0.9
UniModal-LSR (Full)	2025	33.2	38.7	0.8

Table 5: Sign language recognition results.

Method	PHOENIX-14 (WER%)	CSL (Acc%)	WLASL (Acc%)
CNN + LSTM + HMM [40]	26.0	91.2	–
STMC [41]	21.1	94.6	–
FCN [42]	23.3	93.1	–
VAC [43]	21.2	95.2	65.8
UniModal-LSR (Sign only)	19.8	95.8	68.4
UniModal-LSR (Full)	18.3	96.5	70.2
Relative Improvement	13.7%	1.4%	6.7%

Table 6 presents BLEU scores on PHOENIX-2014T. Our model achieves 24.89% BLEU-4, representing an improvement of 2.25 points over the previous state of the art.

Table 6: Sign language translation results on PHOENIX-2014T.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Sign2Text [1]	44.13	31.47	23.89	19.26
Sign2(G+T) [44]	47.26	34.40	26.31	21.32
TSPNet [45]	48.32	35.89	28.12	22.64
UniModal-LSR	51.24	38.67	30.45	24.89
Improvement	+2.92	+2.78	+2.33	+2.25

Table 7 analyzes each component's contribution through cumulative addition, starting from a single-modality baseline and progressively adding each proposed component.

Table 8 provides detailed analysis of CMAF design choices, examining the impact of attention directionality and gating mechanisms.

Table 7: Ablation study: component contributions (WER%).

Configuration	LRS2	PH-14
Single modality	37.9	21.2
+ HTSE	35.8 (−2.1)	20.1 (−1.1)
+ GNN	–	19.4 (−0.7)
+ CMAF	34.2 (−1.6)	18.8 (−0.6)
+ CSA loss	33.2 (−1.0)	18.3 (−0.5)

Table 8: CMAF structural ablation on LRS2.

CMAF Variant	WER (%)	Δ
No cross-modal fusion (concat only)	35.8	–
Unidirectional: Lip→Sign only	34.8	−1.0
Unidirectional: Sign→Lip only	34.6	−1.2
Bidirectional (no gating)	34.1	−1.7
Bidirectional + static gating ($g = 0.5$)	33.8	−2.0
Bidirectional + learned scalar gating	33.5	−2.3
Bidirectional + adaptive element-wise gating	33.2	−2.6

The results confirm several design principles. Bidirectional attention outperforms unidirectional variants because both modalities contain complementary information that benefits the other. Sign-to-lip transfer provides slightly larger gains than lip-to-sign transfer, likely because sign language's richer spatial vocabulary provides additional discriminative cues for disambiguating visually similar lip movements. Adaptive element-wise gating achieves the best performance by allowing dimension-specific modality weighting, enabling the model to selectively combine different aspects of each modality's representation.

Table 9 shows sensitivity to the CSA loss weight λ_{CSA}, and Fig. 4 visualizes the cross-modal alignment effect using t-SNE projections of the learned embeddings.

The visualizations demonstrate that without CSA loss, lip and sign embeddings occupy distinct regions of the representation space, limiting cross-modal transfer. With the CSA loss at $\lambda_{\text{CSA}} = 0.1$, semantically matched pairs from both modalities cluster together, enabling more effective information sharing. The optimal weight balances alignment strength against task-specific discriminability; larger weights improve alignment scores but begin to degrade recognition performance as the representations become overly constrained.

Table 9: Sensitivity analysis of the CSA loss weight (λ_{CSA}). Word error rate (WER%) is reported for LRS2 and PH-14 datasets, while alignment score measures cross-modal alignment quality.

λ_{CSA}	LRS2	PH-14	Alignment Score
0.00	34.8	19.2	0.62
0.05	33.9	18.7	0.71
0.10	33.2	18.3	0.78
0.20	33.5	18.5	0.81
0.50	34.1	19.0	0.84

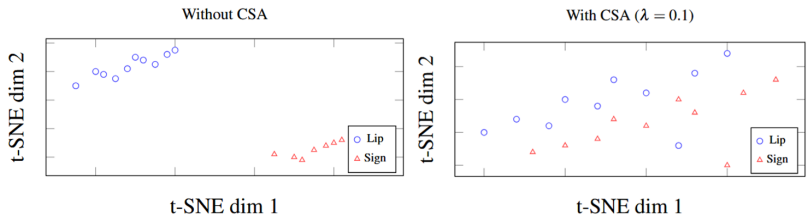

Figure 4: t-SNE visualization of lip and sign embeddings for semantically matched pairs. Without CSA (left), modalities cluster separately in distinct regions. With CSA (right), semantically matched pairs align in the shared embedding space.

Table 10 evaluates performance when one modality is degraded or missing during inference, addressing concerns about robustness to partial input availability.

Table 10: Robustness to modality degradation (inference only).

Condition	LRS2 (WER%)	PH-14 (WER%)	Relative Drop
Full model (both modalities)	33.2	18.3	–
Lip stream only (sign zeroed)	35.1	20.8	+5.7%/+13.7%
Sign stream only (lip zeroed)	36.8	18.9	+10.8%/+3.3%
Lip with 50% frame dropout	34.2	19.5	+3.0%/+6.6%
Sign with hand occlusion (30%)	34.8	19.2	+4.8%/+4.9%
Blurred mouth region (σ = 5)	34.5	19.8	+3.9%/+8.2%

The model exhibits graceful degradation when one modality is impaired. Performance drops range from 3% to 14% depending on the degradation type and target task, demonstrating that cross-modal training provides implicit robustness. As expected, lip reading performance depends more heavily on the lip stream, while sign recognition depends more on the sign stream. Partial degradations such as frame dropout or occlusion have smaller effects than complete modality removal, indicating that the model can leverage whatever partial information remains available. Table 11 compares different pre-training strategies, confirming that joint multimodal pre-training provides the strongest performance gains.

Table 11: Effect of cross-modal pre-training strategies (WER%).

Pre-training Strategy	LRS2	PH-14
None (scratch)	37.9	21.2
Lip only	35.1	20.4
Sign only	36.8	19.6
Joint multimodal	33.2	18.3

Table 12 evaluates generalization by training on one dataset and testing on another without fine-tuning.

The unified model shows improved cross-domain generalization, particularly for lip reading where the 8–9% WER reduction suggests that cross-modal learning provides regularization benefits that transfer across dataset boundaries. The improvement on sign language transfer is smaller but still positive, indicating that the learned representations capture generalizable visual speech features rather than dataset-specific patterns.

Table 12: Cross-domain generalization (Train→Test).

Configuration	LRS2→LRS3	LRS3→LRS2	PH-14→CSL
Single-task	52.3	48.7	82.4
UniModal-LSR	47.8	44.2	85.1
Improvement	−8.6%	−9.2%	+3.3%

Table 13 demonstrates performance under reduced training data availability, showing relative improvements of up to 19% when training with only 10% of the data.

Table 13: Data efficiency: performance vs. training data fraction.

Data %	LRS2 (WER%)		PH-14 (WER%)	
	Baseline	UniModal-LSR	Baseline	UniModal-LSR
10%	58.4	48.2 (−17.5%)	35.6	28.9 (−18.8%)
25%	48.7	41.3 (−15.2%)	28.4	23.5 (−17.3%)
50%	42.1	36.8 (−12.6%)	24.2	20.4 (−15.7%)
100%	37.9	33.2 (−12.4%)	21.2	18.3 (−13.7%)

Table 14 compares computational requirements between separate single-task models and the unified architecture. Latency measurements are end-to-end, including all preprocessing steps, conducted on a single NVIDIA A100 GPU with batch size 1.

Table 14: Computational efficiency comparison.

Method	Params (M)	FLOPs (G)	Latency (ms)	Memory (GB)
Separate (Lip + Sign)	120.6	42.9	98	10.0
UniModal-LSR (Unified)	89.4	31.2	68	7.2
Reduction	−25.9%	−27.3%	−30.6%	−28.0%

We address practical deployment scenarios beyond the full multimodal inference setting. When only one modality is available, such as lip-only or sign-only video, the model can operate with a single branch while still benefiting from joint training. Table 15 shows that single-branch variants extracted from the unified model outperform separately trained single-task models, demonstrating that the cross-modal training provides transferable improvements even when inference uses only one modality. Table 16 presents performance of reduced-capacity variants suitable for edge deployment with limited computational resources.

Table 17 evaluates performance under reduced temporal sampling, which is relevant for bandwidth-constrained applications where transmitting full frame-rate video is impractical.

Fig. 5 decomposes errors by category. The unified model reduces errors across all categories, with notable improvements in viseme confusion (28% reduction) and boundary detection (33% reduction), which are areas where cross-modal information provides the greatest benefit.

Table 15: Single-branch inference performance.

Configuration	LRS2 (WER%)	PH-14 (WER%)	Params (M)
Separate lip-only model	37.9	–	58.2
UniModal-LSR lip branch only	35.6	–	52.1
Separate sign-only model	–	21.2	62.4
UniModal-LSR sign branch only	–	19.8	56.3

Table 16: Lightweight model variants.

Variant	LRS2 (WER%)	PH-14 (WER%)	Params (M)	Latency (ms)
Full model	33.2	18.3	89.4	68
$d = 256$ (half width)	35.8	19.9	45.2	42
$L = 2$ (2 HTSE levels)	34.5	19.2	67.8	51
MobileNet3D backbone	36.2	20.4	32.1	28

Table 17: Performance vs. frame rate.

Frame Rate (fps)	LRS2 (WER%)	PH-14 (WER%)	Latency (ms)
25 (full)	33.2	18.3	68
15	34.1	18.9	45
10	35.8	20.2	32
5	39.4	23.8	18

Figure 5: Error analysis by category showing reductions across all error types.

7 Discussion

Our work builds on the foundation established by multimodal architectures for speech-related tasks. The dual-tower framework proposed by Ge et al. [26] demonstrates the effectiveness of unified multimodal modeling for audio-text fusion in air traffic control communications, achieving improved recognition accuracy through coordinated processing of acoustic and textual information. While their approach improves multimodal speech recognition, it addresses a fundamentally different modality combination. Our CMAF module addresses the unique challenges of visual-visual fusion, where both modalities share temporal structure but differ in spatial semantics. The bidirectional cross-attention design enables more flexible information exchange than parallel tower processing, which is important when the usefulness of lip and

sign cues varies across contexts. Furthermore, the adaptive gating mechanism provides context-dependent weighting that is especially suited to the dynamic relationship between visual speech modalities.

Despite encouraging results, several limitations remain. The benchmark coverage may not fully capture real-world variability in illumination, camera pose, and occlusion, and evaluation on more diverse conditions would strengthen practical applicability. The model is language-specific and must be retrained for each target spoken or signed language, highlighting the need for multilingual or language-agnostic extensions. Training is computationally expensive, requiring eight A100 GPUs during 72 h; although lightweight variants reduce inference cost, they do not address training efficiency. The framework also assumes both lip and sign modalities are available during inference, and while performance degrades gracefully when one modality is missing, explicit strategies such as modality dropout could improve robustness. Finally, the CSA loss relies on semantically aligned lip-sign pairs, and constructing such pairs across datasets may introduce noise, suggesting that future work could explore self-supervised alignment methods that enable training on unpaired data.

8 Conclusion

We introduced UniModal-LSR, a unified multimodal architecture for joint lip reading and sign language recognition. Through hierarchical temporal-spatial encoding, graph-enhanced hand modeling, cross-modal attention fusion with adaptive gating, and contrastive semantic alignment, the model achieves state-of-the-art performance while reducing parameters by 25.9% relative to separate task-specific models. Detailed ablations confirm the contribution of each component, and analysis demonstrates robustness to modality degradation and improved cross-domain generalization.

Future work will explore integration of additional modalities such as audio and full-body pose to capture a more complete picture of visual communication. Language-agnostic representation learning could enable a single model to serve multiple linguistic communities. Self-supervised pre-training methods may reduce labeled data requirements, making the technology more accessible for under-resourced languages. Integration with dialogue systems, such as the full-duplex speech dialogue schemes based on large language models, could enable more natural and responsive human-computer interaction.

Acknowledgement: Not applicable.

Funding Statement: This work is funded by Ho Chi Minh City Open University (HCMCOU) and the Ministry of Education and Training (Vietnam) under grant number B2025-MBS-01.

Author Contributions: The authors report that their specific contributions to this paper are as follows: the study was conceived and designed by Vinh Truong Hoang and Nghia Dinh; data were collected by Thien Ho Huong and Luu Quang Phuong; data analysis and interpretation of the results were undertaken by Kiet Tran-Trung and Ha Duong Thi Hong; and the initial manuscript draft was written by Bay Nguyen Van and Hau Nguyen Trung. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The authors confirm that the data supporting the findings of this study are openly available. LRS2 and LRS3 datasets are available at https://www.robots.ox.ac.uk/vgg/data/lip_reading/. PHOENIX dataset is available at https://www-i6.informatik.rwth-aachen.de/koller/RWTH-PHOENIX/. WLASL dataset is available at https://www.kaggle.com/datasets/risangbaskoro/wlasl-processed. CSL dataset is available at https://ustc-slr.github.io/datasets/2015_csl. How2Sign dataset is available at https://how2sign.github.io/.

Ethics Approval: This research uses publicly available benchmark datasets collected with appropriate consent and ethical approval by the original creators.

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

References

1. Camgoz NC, Hadfield S, Koller O, Ney H, Bowden R. Neural sign language translation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. p. 7784–93. doi:10.1109/CVPR.2018.00812.
2. Assael YM, Shillingford B, Whiteson S, De Freitas N. LipNet: end-to-end sentence-level lipreading. arXiv:1611.01599. 2016.
3. Bear HL, Harvey R. Phoneme-to-viseme mappings: the good, the bad, and the ugly. Speech Commun. 2017;95(3):40–67. doi:10.1016/j.specom.2017.07.001.
4. Heracleous P, Beautemps D, Aboutabit N. Cued speech automatic recognition in normal-hearing and deaf subjects. Speech Commun. 2010;52(6):504–12. doi:10.1016/j.specom.2010.03.001.
5. Petridis S, Stafylakis T, Ma P, Cai F, Tzimiropoulos G, Pantic M. End-to-end audiovisual speech recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2018 Apr 15–20; Calgary, AB, Canada. p. 6548–52. doi:10.1109/ICASSP.2018.8461326.
6. Subba Rao MV, Naga Amulya T, Aparna Y, Pranavi R, Madhumitha S, Priya SS. Speech reconstruction from silent lip movements using deep learning. In: 2025 2nd International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI); 2025 Dec 4–5; Raipur, India. p. 1–6. doi:10.1109/icaiihi67124.2025.11403769.
7. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc.; 2017. p. 6000–10.
8. Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the kinetics dataset. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. p. 4724–33. doi:10.1109/CVPR.2017.502.
9. Yan S, Xiong Y, Lin D. Spatial temporal graph convolutional networks for skeleton-based action recognition. Proc AAAI Conf Artif Intell. 2018;32(1):7444–52. doi:10.1609/aaai.v32i1.12328.
10. van den Oord A, Li Y, Vinyals O. Representation learning with contrastive predictive coding. arXiv:1807.03748. 2018.
11. Potamianos G, Neti C, Gravier G, Garg A, Senior AW. Recent advances in the automatic recognition of audiovisual speech. Proc IEEE. 2003;91(9):1306–26. doi:10.1109/JPROC.2003.817150.
12. Matthews I, Cootes TF, Bangham JA, Cox S, Harvey R. Extraction of visual features for lipreading. IEEE Trans Pattern Anal Mach Intell. 2002;24(2):198–213. doi:10.1109/34.982900.
13. Gales M, Young S. The application of hidden Markov models in speech recognition. Found Trends® Signal Process. 2008;1(3):195–304. doi:10.1561/2000000004.
14. Chung JS, Senior A, Vinyals O, Zisserman A. Lip reading sentences in the wild. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. p. 3444–53. doi:10.1109/CVPR.2017.367.
15. Ma P, Petridis S, Pantic M. End-to-end audio-visual speech recognition with conformers. In: ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2021 Jun 6–11; Toronto, ON, Canada. p. 7613–7. doi:10.1109/ICASSP39728.2021.9414567.
16. Ma P, Petridis S, Pantic M. Visual speech recognition for multiple languages in the wild. Nat Mach Intell. 2022;4(11):930–9. doi:10.1038/s42256-022-00550-z.
17. Shi B, Hsu WN, Lakhotia K, Mohamed A. Learning audio-visual speech representation by masked multimodal cluster prediction. arXiv:2201.02184. 2022.
18. Ma P, Haliassos A, Fernandez-Lopez A, Chen H, Petridis S, Pantic M. Auto-AVSR: audio-visual speech recognition with automatic labels. In: ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2023 Jun 4–10; Rhodes Island, Greece. p. 1–5. doi:10.1109/ICASSP49357.2023.10096889.

19. Shukla A, Petridis S, Pantic M. Does visual self-supervision improve learning of speech representations for emotion recognition? IEEE Trans Affect Comput. 2023;14(1):406–20. doi:10.1109/TAFFC.2021.3062406.
20. Prajwal KR, Afouras T, Zisserman A. Sub-word level lip reading with visual attention. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 5152–62.
21. Hu H, Zhao W, Zhou W, Wang Y, Li H. SignBERT: pre-training of hand-model-aware representation for sign language recognition. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. p. 11067–76. doi:10.1109/ICCV48922.2021.01090.
22. Graves A, Fernández S, Gomez F, Schmidhuber J. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the 23rd International Conference on Machine Learning—ICML '06; 2006 Jun 25–29; Pittsburgh, PA, USA. p. 369–76. doi:10.1145/1143844.1143891.
23. Lu J, Yang J, Batra D, Parikh D. Hierarchical question-image co-attention for visual question answering. arXiv:1606.00061. 2016.
24. Tsai YH, Bai S, Liang PP, Kolter JZ, Morency LP, Salakhutdinov R. Multimodal transformer for unaligned multimodal language sequences. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL; 2019. p. 6558–69. doi:10.18653/v1/p19-1656.
25. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. arXiv:2103.00020. 2021.
26. Ge S, Ren J, Shi Y, Zhang Y, Yang S, Yang J. Audio-text multimodal speech recognition via dual-tower architecture for mandarin air traffic control communications. Comput Mater Contin. 2024;78(3):3215–45. doi:10.32604/cmc.2023.046746.
27. Li G, Deng J, Geng M, Jin Z, Wang T, Hu S, et al. Audio-visual end-to-end multi-channel speech separation, dereverberation and recognition. IEEE/ACM Trans Audio Speech Lang Process. 2023;31:2707–23. doi:10.1109/TASLP.2023.3294705.
28. Wang X, Jiang H, Huang H, Fang Y, Xu M, Wang Q. DCIM-AVSR: efficient audio-visual speech recognition via dual conformer interaction module. In: ICASSP 2025–2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2025 Apr 6–11; Hyderabad, India. p. 1–5. doi:10.1109/ICASSP49660.2025.10890272.
29. Zhang F, Bazarevsky V, Vakunov A, Tkachenka A, Sung G, Chang C-L, et al. MediaPipe hands: on-device real-time hand tracking. arXiv:2006.10214. 2020.
30. Yun C, Bhojanapalli S, Rawat AS, Reddi S, Kumar S. Are transformers universal approximators of sequence-to-sequence functions? arXiv:1912.10077. 2020.
31. Afouras T, Chung JS, Zisserman A. LRS3-TED: a large-scale dataset for visual speech recognition. arXiv:1809.00496. 2018.
32. Cooke 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 Pt 1):2421–4. doi:10.1121/1.2229005.
33. Koller O, Forster J, Ney H. Continuous sign language recognition: towards large vocabulary statistical recognition systems handling multiple signers. Comput Vis Image Underst. 2015;141(5):108–25. doi:10.1016/j.cviu.2015.09.013.
34. Huang J, Zhou W, Zhang Q, Li H, Li W. Video-based sign language recognition without temporal segmentation. Proc AAAI Conf Artif Intell. 2018;32(1):2257–64. doi:10.1609/aaai.v32i1.11903.
35. Li D, Opazo CR, Yu X, Li H. Word-level deep sign language recognition from video: a new large-scale dataset and methods comparison. In: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV); 2020 Mar 1–5; Snowmass Village, CO, USA. p. 1448–58. doi:10.1109/wacv45572.2020.9093512.
36. Duarte A, Palaskar S, Ventura L, Ghadiyaram D, DeHaan K, Metze F, et al. How2Sign: a large-scale multimodal dataset for continuous American sign language. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA. p. 2734–43. doi:10.1109/cvpr46437.2021.00276.
37. Afouras T, Chung JS, Senior A, Vinyals O, Zisserman A. Deep audio-visual speech recognition. IEEE Trans Pattern Anal Mach Intell. 2022;44(12):8717–27. doi:10.1109/TPAMI.2018.2889052.

38. Martinez B, Ma P, Petridis S, Pantic M. Lipreading using temporal convolutional networks. In: ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2020 May 4–8; Barcelona, Spain. p. 6319–23. doi:10.1109/icassp40776.2020.9053841.
39. Kim M, Yeo JH, Ro YM. Distinguishing homophenes using multi-head visual-audio memory for lip reading. Proc AAAI Conf Artif Intell. 2022;36(1):1174–82. doi:10.1609/aaai.v36i1.20003.
40. Koller O, Camgoz NC, Ney H, Bowden R. Weakly supervised learning with multi-stream CNN-LSTM-HMMs to discover sequential parallelism in sign language videos. IEEE Trans Pattern Anal Mach Intell. 2020;42(9):2306–20. doi:10.1109/TPAMI.2019.2911077.
41. Zhou H, Zhou W, Zhou Y, Li H. Spatial-temporal multi-cue network for continuous sign language recognition. Proc AAAI Conf Artif Intell. 2020;34(7):13009–16. doi:10.1609/aaai.v34i07.7001.
42. Cheng KL, Yang Z, Chen Q, Tai YW. Fully convolutional networks for continuous sign language recognition. In: Computer Vision—ECCV 2020. Cham, Switzerland: Springer; 2020. p. 697–714. doi:10.1007/978-3-030-58586-0_41.
43. Min Y, Hao A, Chai X, Chen X. Visual alignment constraint for continuous sign language recognition. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. p. 11522–31. doi:10.1109/ICCV48922.2021.01134.
44. Cihan Camgoz N, Koller O, Hadfield S, Bowden R. Sign language transformers: joint end-to-end sign language recognition and translation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 10020–30. doi:10.1109/cvpr42600.2020.01004.
45. Li D, Xu C, Yu X, Zhang K, Swift B, Suominen H, et al. TSPNet: hierarchical feature learning via temporal semantic pyramid for sign language translation. In: NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc.; 2020. p. 12034–45.