LinProVSR: Linguistics-Knowledge Guided Progressive Disambiguation Network for Visual Speech Recognition

Feng Xue[1]*, Baochao Zhu[2], Wei Jia[2], Shujie Li[1], Yu Li[2], Jinrui Zhang[2], Shengeng Tang[2], Dan Guo[2]

[1]School of Software, Hefei University of Technology
[2]School of Computer Science and Information Engineering, Hefei University of Technology
feng.xue@hfut.edu.cn,baochao.zhu@mail.hfut.edu.cn,jiawei@hfut.edu.cn,lisjhfut@hfut.edu.cn
yuli@mail.hfut.edu.cn,2024010064@mail.hfut.edu.cn,tangsg@hfut.edu.cn,guodan@hfut.edu.cn

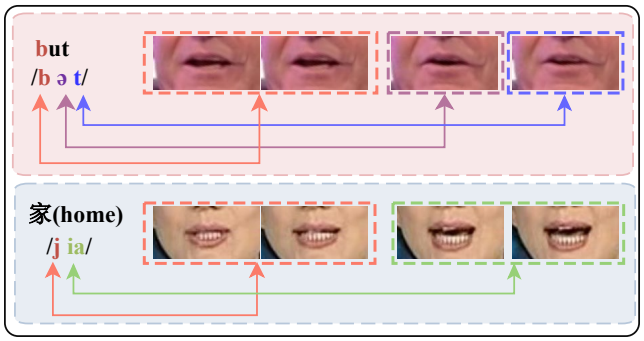
Abstract

Visual Speech Recognition (VSR), commonly known as lipreading, enables the recognition of spoken text by analyzing lip visual features. Due to the subtlety of lip movements, its recognition is much harder than other motion recognition tasks. Existing VSR models face the challenge of viseme ambiguity when processing phonemes with similar pronunciations—multiple phonemes share similar viseme features, leading to a notable drop in lipreading accuracy. To address this issue, this study proposes a Linguistics-Knowledge Guided Progressive Disambiguation Network for Visual Speech Recognition(**LinProVSR**) framework. First, an ambiguous sample set is constructed based on linguistic knowledge to provide supervisory signals for the model's training. Then, a Progressive Contrastive Disambiguation Network (PCDN) is designed, which progressively enhances the model's ability to capture the subtle viseme differences corresponding to similar phonemes through viseme-phoneme contrastive disambiguation in the encoding stage and text contrastive disambiguation in the decoding stage. Furthermore, we pioneer the Ambiguous Word Error Rate (AWER) metric specifically for evaluating recognition of phonetically ambiguous text, and verify the effectiveness of the proposed method on multiple public datasets, achieving a significant breakthrough especially in distinguishing visually similar phonemes.

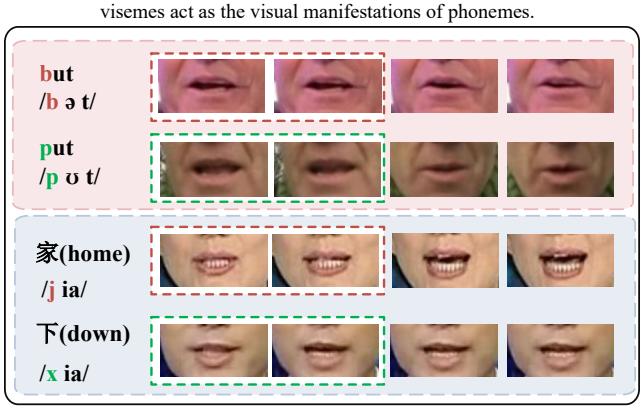
Introduction

Visual Speech Recognition (VSR) not only builds a communication bridge for the hearing-impaired community but also demonstrates its indispensable application potential in urban public safety and scenarios with missing audio(Zhang et al. 2024; Yeo et al. 2024; Song et al. 2024; Ma et al. 2022; Cheng et al. 2023; Xu et al. 2020; Bai et al. 2021; Kumar et al. 2020).

Unlike gesture and motion recognition tasks(Siddiqui, Tirupattur, and Shah 2024; Zhang et al. 2025; Do and Kim 2025; Guo et al. 2024), lip movements exhibit highly subtle and dynamically complex characteristics. Minuscule muscle movements of the lips—such as the degree of lip opening and closing, and the hidden movements of the tongue tip—often carry crucial articulatory information. In VSR, the primary challenge stems from viseme ambiguity: multiple phonemes with distinct pronunciations often share identical or highly similar visemes, causing visual signals to fail in uniquely mapping to phonemes and thereby leading to recognition errors.(Hao et al. 2025; Harte and Gillen 2015; Zhou et al. 2014). As illustrated in Fig. 1: (a) Both English and Chinese pronunciations can be broken down into minimal units of sound—phonemes(e.g., /b/,/ə/,/t/ in 'but') These phonemes visually correspond to sequences of video frames of varying lengths, referred to as visemes. Importantly, the mapping between phonemes and visemes is not strictly one-to-one; (b) Illustration of viseme ambiguity: phonemes with



(a) Illustration of the mapping between visemes and phonemes, where visemes act as the visual manifestations of phonemes.



(b) Illustration of viseme ambiguity. Visually similar phonemes(viseme) leads to recognition ambiguity.

Figure 1: Illustration of the visemes-phonemes ambiguity.

*Corresponding author.

similar pronunciations exhibit highly similar viseme representations, and this visual confusion directly introduces ambiguity into the recognition process. (e.g., due to similar visemes, 'but' is often misrecognized as 'put').

To address this challenge, current VSR methods mainly use two strategies: enhancing visual encoders(Chung et al. 2017a; Xu et al. 2018; Zhang et al. 2019; Xue et al. 2023a) and multimodal fusion(Weng and Kitani 2019; Zhao et al. 2020; Li et al. 2024). Visual encoder enhancement focuses on designing more powerful architectures to better capture discriminative features from lipreading videos. However, their effectiveness is limited by insufficient visual semantics and lack of explicit supervision for ambiguous phonemes, making it hard to reliably distinguish similar lip configurations. Multimodal fusion methods integrate complementary signals like facial landmarks and optical flow to go beyond unimodal vision limitations. They enhance visual feature separability through cross-modal alignment during training but lack a deep semantic model for resolving ambiguous visual representations. Guided by the linguistic insight of asymmetric phoneme-viseme mapping, We propose the LinProVSR framework. It deeply integrates linguistic knowledge priros into lipreading feature learning and decoding, establishing a dynamically adaptable ambiguity resolution mechanism, consisting of two core modules:

(1) Language-prior-based Ambiguous Sample Construction Module (ASCM). This module automatically replaces the text labels of training samples with corresponding ambiguous texts according to linguistic pronunciation rules. These are not simple data augmentation methods, but rather interpretable ambiguous sample replacement strategies, which provide precise training targets for subsequent ambiguity recognition.

(2)The Progressive Contrastive Disambiguation Network (PCDN) uses a two-stage, closed-loop process for cross-modal ambiguity resolution. Stage A (Encoding): A triplet-based contrastive loss aligns visual viseme features with ground-truth and ambiguous phoneme features, forcing the encoder to learn discriminative articulatory cues. Stage B (Decoding): A text semantic contrastive loss constrains predictions against ground-truth and ambiguous texts. This leverages linguistic knowledge for corrections, creating a feature refinement loop that enables end-to-end disambiguation. In Stage A, the contrastive learning among visemes, ground-truth phonemes, and ambiguous phonemes constitutes the core and primary mechanism of LinProVSR's disambiguation capability, while Stage B serves as a supplementary error-correction mechanism to address residual errors that evaded detection in Stage A.

Our contributions are summarized as follows.

- Based on the refined phonological rules of speech articulatory movements, we propose a linguistically oriented ambiguous sample construction strategy to provide explicit supervision for downstream disambiguation of viseme-phoneme asymmetric ambiguities.
- We innovatively design a progressive cross-modal contrastive disambiguation network(closed-loop paradigm), integrating dynamic visual-phoneme mapping in encoding and enables semantic self-correction in decoding to achieve end-to-end ambiguity resolution from initial features to final semantics.
- We for the first time proposed the AWER metric for evaluating recognition of phonetically ambiguous text; extensive validation on public datasets confirmed our method's effectiveness, especially in markedly improving the distinction of ambiguous visemes.

Related Work

Early research on Visual Speech Recognition (VSR) focused on extracting static lip shape features, predominantly utilizing CNNs to capture lip textures. However, such methods overlooked the temporal dynamics of lip movements and struggled to model lip shape variation patterns during phoneme articulation. To address this limitation, researchers began exploring joint modeling of spatiotemporal features, introducing 3DCNNs to extract spatiotemporal features from lipreading videos, combined with GRU and CTC for spoken text decoding (Assael et al. 2016; Xu et al. 2018; Son and Zisserman 2017). As the demand for modeling long-range temporal dependencies in VSR tasks became increasingly prominent and with the widespread adoption of Transformer models in sequence modeling tasks, studies introduced Transformer architectures into VSR tasks to design lipreading network structures(Prajwal, Afouras, and Zisserman 2022; Park, Park, and Park 2025; Wang et al. 2024; Kit Khinn Teng, Zhang, and Saitoh 2025).

Additionally, Xue, Li, and others incorporated facial landmarks to construct multimodal fusion mechanisms, alleviating visual ambiguity by supplementing contextual information(Xue et al. 2023a; Li et al. 2024). In contrast, the progressive contrastive disambiguation model proposed in this paper, by deeply integrating linguistic knowledge into the processes of feature learning and spoken text decoding, specifically addresses the viseme-phoneme asymmetric ambiguity and thus exhibits unique advantages in distinguishing similar lip shapes.

The CALLip model(Huang, Liang, and Fang 2021) Enhances the discriminability of visual features through attribute learning and cross-modal contrastive learning; Wang et al. (Wang et al. 2023) proposed the TalkLip model, which utilizes contrastive learning to improve lipreading synchronization. In contrast to these methods, the progressive contrastive cross-modal disambiguation network proposed in this paper is not merely a simple comparison between positive and negative samples. Instead, it achieves targeted resolution of viseme-semantic ambiguities through a closed-loop logic of interpretable sample construction-dynamic mapping-semantic error correction. This mechanism is fundamentally different from pure contrastive learning.

Methodology

The architecture of our LinProVSR model is illustrated in Fig. 2, consisting of two core modules: (1) Linguistic prior-driven Ambiguity Sample Construction Module (ASCM): Based on the viseme-phoneme mapping rules in linguistics, ASCM constructs a set of ambiguous texts that have similar

Language	phoneme Num	phoneme Group	Viseme Num
English	48	(1) ow,oy (2) b,p,m ... (18) d,t,n,l	18
Chinese	32	(1) b,p,m (2) d,t,n,l ... (12) z,c,s	12

Table 1: Linguistic Knowledge and Viseme Details

lip shapes to the ground truth texts but different semantics by replacing similar phonemes in the pronunciation texts. (2) Progressive Contrastive Disambiguation Network (PCDN): This module is composed of a viseme-phoneme disambiguation sub-module and a text semantic disambiguation sub-module. Through viseme-phoneme feature alignment and intra-modal text semantic alignment, these two disambiguation sub-modules form a closed-loop disambiguation process from feature optimization to semantic error correction.

Linguistic Prior-Driven Ambiguous Sample Construction Module(ASCM)

In human linguistics, there are three fundamental consensual knowledge rules(as shown in Table 1):

(1) Human speech (both in English and Chinese) can be segmented into minimal pronunciation units (phonemes). Among them, English has approximately 48 phonemes, while Mandarin Chinese has about 32 phonemes.

(2) Several phonemes that have similar lip shapes (with subtle differences) during pronunciation can be classified into the same viseme.

(3) The ambiguity of English visemes mainly originates from consonants and some vowels (such as /iː/ and /ɪ/); the ambiguity of Chinese visemes mainly comes from initial consonants (such as /z/ and /c/) and tones.

We first convert words from datasets into phoneme sequences using the CMU Pronouncing Dictionary and Modern Chinese Dictionary. Subsequently, phonemes within the same viseme group are randomly replaced to generate corresponding ambiguous texts. For example,we can replace 'sip (/s/)' with 'zip (/z/)' and 'da1' with 'ta2'.

$$Text_{amb} = Text_{original}[i] + k, \tag{1}$$

where i is the character position in the original text, and k is the number of tones to change.

Thus, the ambiguous text generation strategy based on linguistic knowledge can be expressed as follows:

$$Text_{amb} = Gen(Text_{original}, Rule_{replace}), \tag{2}$$

where $Text_{amb}$ represents the generated ambiguous text, $Text_{original}$ represents the original input text, $Gen(\cdot)$ is a generation function that generates the ambiguous text based on specific transformation rules, $Rule_{replace}$ is a set of rules defining how elements in the input text are substituted to generate the ambiguous text.

Progressive Contrastive Disambiguation Network(PCDN)

(1) Viseme-Phoneme Disambiguation Module

Textual-Visual Feature Encoding. During the encoding stage, our objective is to accurately map viseme-phoneme while boosting the discriminative capacity of visual features. We begin by representing the lipreading video as $x \in \mathbb{R}^{T\times H\times W\times 3}$, where T denotes the video duration, and H and W represent the frame height and width, respectively. This video is then fed into a 3D CNN for feature extraction from the lip region. Positional Encoding is added to retain the sequential positional information, resulting in the extraction of its spatiotemporal features, denoted as $F_v \in R^{T\times d}$, where d represents the feature dimension of the video frames:

$$F_v = PE(3DCNN(x)), \tag{3}$$

where $PE(\cdot)$ represents the positional encoding.

Words from $Text_{original}$ and $Text_{amb}$ are mapped to embedding vectors via a neural network, enhanced with positional encoding to capture sequential context. The resulting representations jointly encode semantic and positional features:

$$F_t = PE(Embedding(Text_{original})) \in \mathbb{R}^{L\times d}, \tag{4}$$

$$F_s = PE(Embedding(Text_{amb})) \in \mathbb{R}^{L\times d}, \tag{5}$$

where L is the length of the text sequence, and d is the size of the word embedding vectors, which matches the dimensionality of the video feature vectors. Specifically, the spatiotemporal visual features F_v are first linearly projected to obtain the query Q_v, key K_v, and value $V_v$representations.

$$Q_v = F_v W_Q^v, K_v = F_v W_K^v, V_v = F_v W_V^v, \tag{6}$$

where W_Q^v, W_K^v, W_V^v are the learnable weight matrices.

The video sequence encoder employs an attention mechanism by computing weights between its query vectors Q_v and key vectors K_v. Formally, this process is articulated as:

$$f_{att}^v = Softmax(\frac{Q_v K_v^T}{\sqrt{d}})V_v, \tag{7}$$

where d represents the dimensionality of the key vector.

Building on the basic attention logic, the video sequence encoder deploys a Multi-Head Attention mechanism to capture features across multiple semantic levels via parallelized computations, dynamically model temporal dependencies among consecutive lipreading video frames with finer granularity, concatenate and project independent attention outputs from each head through a linear layer, apply a residual connection to preserve original feature information while fusing multi-head insights, and ultimately yield multi-scale dynamic temporal features f_{mdt}^v, then enter a feed-forward network to model complex non-linearities. A parallel residual connection enforces temporal coherence in the final viseme feature representation. f_v:

$$f_{ffn}^v = Linear(ReLU(Linear(LN(f_{mdt}^v)))), \tag{8}$$

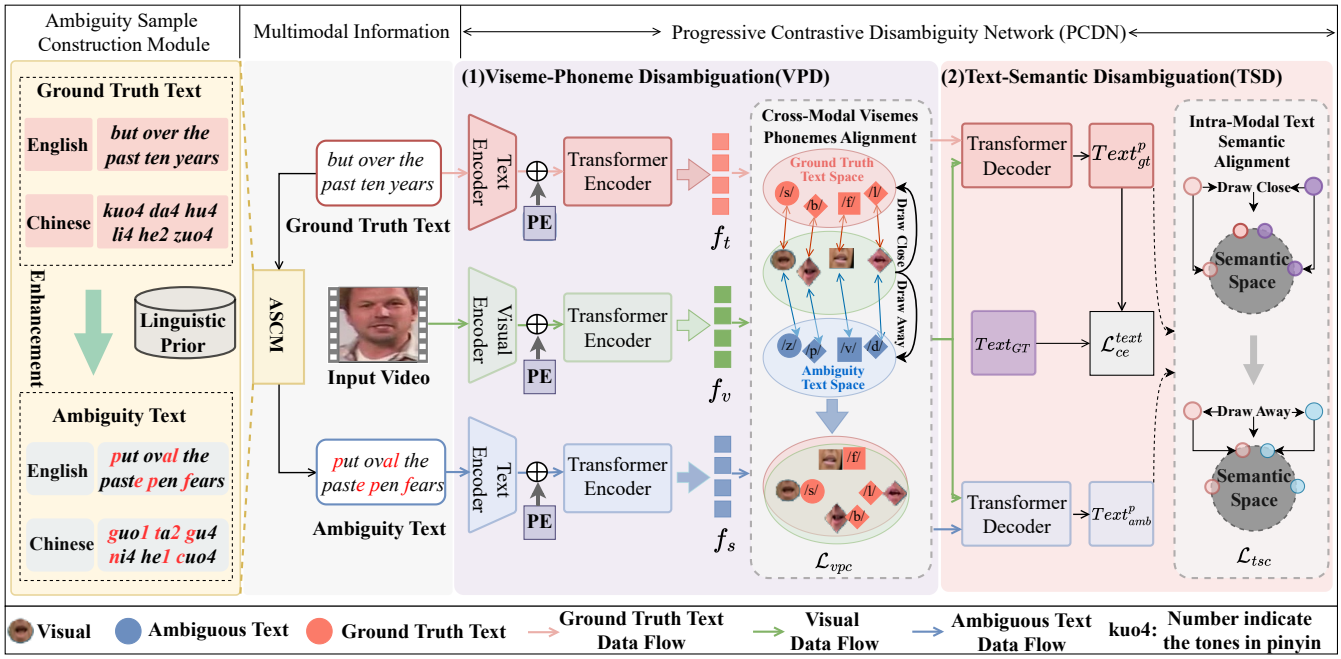


Figure 2: Overview of LinProVSR framework.

where LN is the Layer Normalization.

$$f_v = LN(f^v_{mdt} + f^v_{ffn}) \tag{9}$$

Similarly, a text sequence encoder, sharing the same architecture as the video sequence encoder, is used to extract high-level semantic features from both original and ambiguous texts, while capturing contextual dependencies within their respective phoneme sequences. This process yields the true text features f_t and the ambiguous text features f_s.

Viseme-Phoneme Contrastive Loss. Building upon this foundation, we introduce a viseme-phoneme contrastive loss function $\mathcal{L}_{vpc}$, designed to achieve high-quality mapping between visemes and phonemes, thereby enhancing the discriminative capacity of the visual encoder when encoding visemes associated with ambiguous phonemes. Specifically, this contrastive loss imposes similarity alignment constraints on three types of features: viseme features f_v extracted by the aforementioned encoder, ground-truth textual phoneme features f_t, and ambiguous textual phoneme features f_s. This formulation ensures the aggregation of ground-truth phoneme features with viseme features in the feature space, while simultaneously widening the representational distance between ambiguous phoneme features and viseme features within the same space. By implementing this mechanism, the model learns fine-grained mapping relationships between phonemes and visemes in ground-truth text, captures subtle visual distinctions corresponding to similarly pronounced phonemes, and ultimately improves lipreading accuracy for ambiguously articulated phonemes. We transpose the dimensions of f_t,f_s and f_v, and apply mean pooling along the frame and phoneme sequence dimensions to ensure consistent dimensions.

$$\mathcal{L}_{vpc} = \max\left(0, \frac{\|\mathbf{f}_v - \mathbf{f}_t\|_2^2 - \|\mathbf{f}_v - \mathbf{f}_s\|_2^2}{\tau_0 \cdot e^{-k \cdot t}} + \alpha\right), \tag{10}$$

where α controls the minimum distance difference between positive and negative samples and is typically set between 0.1 and 1.0, τ_0 denote the initial temperature (default value is 0.07), k the decay coefficient (default value is 0.001), t the normalized training step.

(2) Text Semantic Disambiguation Module During the decoding stage, to enable the model to perform ambiguity discrimination by integrating linguistic knowledge, we adopt a visual-text to word decoding approach for sequence decoding. Specifically, Multi-Head Attention is applied to both the ground truth embeddings F_t and the ambiguity text embeddings F_s, yielding representations f^t_{mha} and f^s_{mha}. Multi-Head Cross Attention is then applied to the video encoder output, separately obtaining corresponding weighted contextual representations that correspond to the video sequence attended by the ground truth sequence and the ambiguous text sequence. These weighted contextual representations then pass through a feed-forward neural network, and the outputs f^t_d and f^s_d are obtained after applying residual connections. These outputs are processed through a linear layer and a softmax to obtain the predicted probabilities for the current word. The word with the highest probability is selected as the current prediction $Text^p_{gt}$ and $Text^p_{amb}$. $\text{Text}^p_{gt} \triangleq (p^{gt}_{l,v})_{L\times V}$ represents the predicted word corresponding to the v-th word in the vocabulary based on the ground truth. Similarly, $Text^p_{amb} \triangleq (p^{amb}_{l,v})_{L\times V}$ represents the predicted word corresponding to the v-th word in the vocabulary based on the ambiguous text. The words with the

highest probabilities are selected as the current predictions y_{gt}^{text} and y_{amb}^{text}.

The process repeats until the predicted word is '<eos>' or the predefined text length threshold is reached, yielding the complete predicted character sequences y_{gt}^{text} and y_{amb}^{text}. In contrast, Chinese decoding uses the same decoder architecture. However, since Chinese is a tonal language with a large number of homophones and near-homophones, we adopt a two-stage cascaded decoding strategy: from visual-text to Pinyin, and then from Pinyin to Chinese characters.

Text Semantic Contrastive Loss. To further enhance the LinProVSR model's ability to discern subtle differences in lipreading visual elements corresponding to similar ambiguous phonemes, we designed a contrastive loss termed $\mathcal{L}_{tsc}$ based on triplets $<Text_{gt}^{p}, Text_{amb}^{p}, Text_{gt}>$, where: $Text_{gt}^{p}$ represents the text jointly decoded from original text features and visual features, $Text_{amb}^{p}$ denotes the text derived from ambiguous text features combined with visual features, and $Text_{gt}$ is the ground-truth (GT) text. By imposing constraints on this triplet, we simultaneously enforce semantic consistency between $Text_{gt}^{p}$ and $Text_{gt}$ while amplifying semantic dissimilarity between $Text_{gt}^{p}$ and $Text_{amb}^{p}$. Through backpropagation of the contrastive loss, the model is compelled to trace back and further strengthen its capability to differentiate lip-feature variations associated with these two phonemes.

$$D = \left(d(Text_{gt}, Text_{gt}^{p}) - d(Text_{gt}, Text_{amb}^{p}) + m\right), \tag{11}$$

$$\mathcal{L}_{tsc} = \max\left(D, 0\right), \tag{12}$$

where $d(\cdot)$ represents the Euclidean distance between the two, $Text_{gt}$ is the true Pinyin sequence, and m is a hyper-parameter that defines the minimum allowable distance difference.

Loss Function

Our model utilizes three loss functions during training. The first is the Viseme-Phoneme Contrastive loss, the second is the Text Semantic Contrastive loss, and the third loss is CrossEntropy loss, denoted as $\mathcal{L}_{ce}^{text}$:

$$\mathcal{L}_{ce}^{text} = -\sum_{i=1}^{L} Text_{gt} \log Text_{gt}^{p}, \tag{13}$$

Cross-entropy loss serves to align the model's predictions with the true target text. Through the minimization of such losses, the model is enabled to learn effectively and produce outputs that align with the ground truth.

The overall loss of LinProVSR can be defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{vpc} + \mathcal{L}_{tsc} + \mathcal{L}_{ce}^{text}, \tag{14}$$

For Chinese input, the $\mathcal{L}_{ce}^{text}$ loss is defined as follows:

$$\mathcal{L}_{ce}^{text} = -\sum_{i=1}^{L} (Text_{gt} \log Text_{gt}^{p} + Text_{gt}^{hz} \log Text_{gt}^{p}). \tag{15}$$

Dataset	Vocabulary	Train	Valid	Test
CMLR	3517	71448	10206	20418
LRS2	62769	45839	1082	1243
GRID	51	24750	-	8250

Table 2: Statistical data of datasets CMLR, LRS2 and GRID.

Experiments

Experimental Settings

Dataset. Experiments were conducted on three large-scale lipreading datasets. As shown in Table 2: CMLR(Zhao, Xu, and Song 2019), a Mandarin Chinese dataset, and LRS2(Chung et al. 2017b) and GRID(Cooke et al. 2006), two widely-used English-language datasets.

Evaluation Metrics. Within VSR task, Error Rate is a widely adopted performance metric, as its reduction directly indicates an improvement in overall model performance. To comprehensively assess the model across different linguistic settings, we employ Word Error Rate (WER) as the primary evaluation metric.

Baselines. We compare the performance of our model with that of several mainstream architectures in current VSR research, including CSSMCM(Zhao, Xu, and Song 2019), LIBS(Zhao et al. 2020), CALLip(Huang, Liang, and Fang 2021), LCSNet(Xue et al. 2023b), LipFormer(Xue et al. 2023a), GUSLip(Li et al. 2024), WAS(Chung et al. 2017b), CTC/Att(Petridis et al. 2018), TDNN(Yu et al. 2020), Pan et al(Pan et al. 2022), TM-seq2seq(Afouras et al. 2022), and Fca-Net(Yang, Gong, and Kang 2023), LiteVSR(Laux et al. 2024), LiteVSR2(Laux and Schmeink 2024), SwinLip(Park, Park, and Park 2025), Wu et al(Wu et al. 2024), CFLip(Li et al. 2025). To ensure a fair comparison based solely on visual input, we exclude methods that leverage additional audio information or knowledge distillation techniques from our evaluation.

Implementation Details. We first use the DLib face detector(King 2009) to locate 20 lip key points on the speaker's face in the video. The lip region is then cropped and scaled to 64x128 pixels. During the training phase, we set the batch size to 8 and adopted a learning rate warm-up strategy to stabilize the training and avoid local minima. Specifically, in the Adam optimizer(Adam et al. 2014), the learning rate dynamically adjusts according to the error rate, with a peak value set at 0.0003.

Comparison Experiments

In this experiment, we compare the performance of our method with the baseline across three lipreading datasets.

CMLR and GRID. Table 3 compares LinProVSR with established lipreading baselines, where "-" denotes unavailable metrics. On the CMLR dataset, LinProVSR achieves a 23.34% WER, outperforming all baselines and thus validating its effectiveness on Chinese data. On the GRID dataset,

Method	CMLR WER(%)↓	GRID WER(%)↓
CSSMCM(2019)	32.48	-
LIBS(2020)	31.27	-
CALLip(2021)	31.18	2.48
LCSNet(2023)	30.03	2.30
LipFormer(2023)	27.79	1.45
Wu et.al(2024)	-	1.83
GUSLip(2024)	29.98	1.93
CFLip(2025)	26.20	1.01
LinProVSR	**23.34**	**0.99**

Table 3: Comparison of results between LinProVSR and other baseline models on CMLR and GRID.

Method	Training Datasets Used	WER(%)↓
CTC/Att(2018)	LRS2,LRW	63.5
LIBS(2020)	LRS2,LRS3	65.3
TDNN(2020)	LRS2	48.9
TM-seq2seq(2022)	LRS2,LRS3,LRW MV-LRS	48.3
Pan et.al(2022)	LRS2,LRW	43.2
Fca-Net(2023)	LRS2,LRW	38.3
LiteVSR(2024)	LRS2,LRS3	47.4
LiteVSR2(2024)	LRS2,LRS3	40.6
SwinLip(2024)	LRS2	37.01
LinProVSR	**LRS2**	**35.36**

Table 4: Comparison of results between LinProVSR and other baseline models on LRS2.

it reaches a 0.99% WER, a 1.49% reduction from CALLip, confirming its effectiveness on English data.

LRS2. To further validate the performance of the LinProVSR on large-scale English datasets, we compared it with state-of-the-art baseline models on the widely used dataset LRS2.

Table 4 presents the performance of the LinProVSR model on the large-scale English dataset LRS2. In Table 4, the "Training Datasets Used" column specifies the datasets employed during the model's training process. Notably, LinProVSR achieves a WER of 35.36% on the LRS2 dataset, outperforming both methods trained on equivalent datasets and those leveraging larger training datasets. This underscores its effectiveness in handling large-scale English datasets while requiring fewer training resources.

Ablation Studies

To evaluate the contributions of Viseme-Phoneme Disambiguation(VPD) and Text Semantic Disambiguation(TSD) modules in reducing lipreading error rates, we perform ablation studies on CMLR, GRID, and LRS2 datasets.

#	VPD	TSD	CMLR WER(%)↓	GRID WER(%)↓	LRS2 WER(%)↓
0	-	-	25.90	2.76	43.02
1	-	√	25.64	1.34	40.31
2	√	-	25.23	1.28	36.25
3	√	√	**23.34**	**0.99**	**35.36**

Table 5: Ablation studies for VPD and TSD on CMLR, GRID and LRS2

Table 5 reports the results of ablation experiments on the VPD and TSD modules. Experimental data indicate that the baseline model (#0) attains WER of 25.90%, 2.76%, and 43.02% on the CMLR, GRID, and LRS2 datasets, respectively. Here, *baseline* denotes the LinProVSR architecture without the VPD and TSD components, which is referred to as the base henceforth (used in Table 6, Fig. 3, and Fig. 4). Comparative analysis reveals that Models #1 and #2 exhibit lower WER values than the base across all three datasets, thereby demonstrating the effectiveness of both modules. Notably, Model #3 achieves the lowest WER, signifying that integrating VPD and TSD modules substantially enhances recognition accuracy. Furthermore, Model #2 outperforms Model #1 in performance metrics, underscoring the VPD module's greater contribution to improved recognition performance. This finding corroborates our theoretical framework: VPD serves as the primary mechanism for resolving ambiguous visemes, while TSD functions as a complementary secondary system for disambiguation and error correction.

Case Study

To evaluate the performance of the proposed model, we present the prediction results for a series of cases, where the characters marked in red indicate incorrect predictions.

In Table 6, we compare the prediction differences between the baseline model and our proposed LinProVSR on the CMLR, GRID, and LRS2 datasets, respectively. The baseline model also mispredicts words(e.g., classifies "mo2" as "bo2" and "right" as "white"). In contrast, our proposed LinProVSR accurately recognizes these characters, demonstrating its effectiveness in handling texts with similar pronunciations and lip movements.

Visualization Analysis

In Fig. 3, we compare heatmaps of weight vectors from the visual encoders of the Baseline and LinProVSR. Thanks to VPD and TSD modules' contrastive constraints, LinProVSR forces the visual encoder to focus more on regions discriminative for ambiguous word pronunciation (e.g., lips, mouth), while the Baseline's focus is more scattered.

As shown in Table 3 and Table 4, the LinProVSR outperforms all baseline models. To further verify that the advanced performance of our proposed model mainly benefits

Case	Method	CMLR		GRID		LRS2	
		Predict	WER(%)	Predict	WER(%)	Predict	WER(%)
Case1	Base	**shi4 shi2 yi1 ju4** zhi3 chu1	66.67	lay blue **by z** nine now	33.33	and we **wore white**	50.00
	LinProVSR	shi2 shi1 yi4 jia4n zhi3 chu1	0.00	lay blue in b nine now	0.00	and we were right	0.00
Case2	Base	**jiu3** ye4 gui **bo2 guo4** da4	50.00	set red **at z** five now	33.33	they're **mowing the ground**	75.00
	LinProVSR	jiu4 ye4 gui1 mo2 kuo4 da4	0.00	set red in **z** five now	16.67	they're moving around	0.00

Table 6: Case study on CMLR, GRID and LRS2

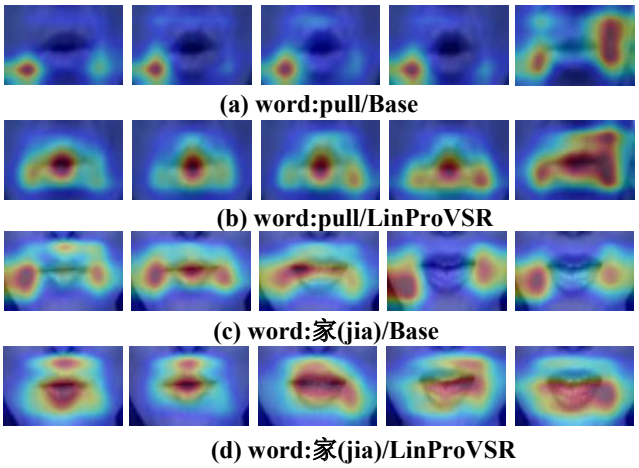
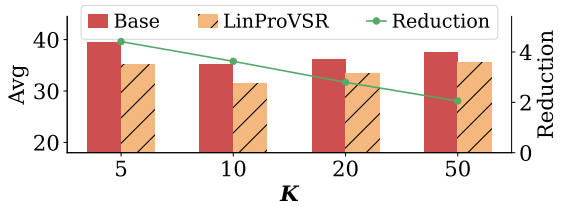
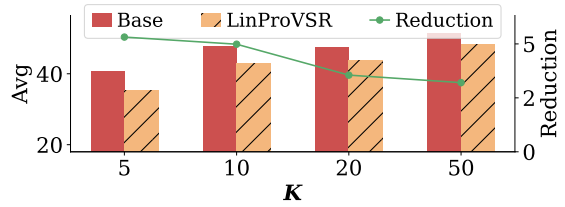


Figure 3: Heatmap comparison between Base and LinProVSR on lip regions



(a) CMLR



(b) LRS2

Figure 4: AWER comparison between Base and LinProVSR

from the high-precision recognition of ambiguous words, we propose for the first time the AWER metric specifically designed to measure the error rate of ambiguous words:

$$AWER(i) = \frac{\text{Num}(\text{pred}(A_i) = \text{ERROR})}{\sum_{n=1}^{N} \text{Num}(\text{pred}(A_n) = \text{ERROR})} \quad (16)$$

Where, A_i represents the i-th ambiguous word in the ambiguous word set, and denominator denotes the total number of erroneous predictions for A_i across the test set. The denominator aggregates the total error counts for all ambiguous words. Given potential variability in individual ambiguous word error rates, we employ averaged AWER to rigorously assess the effectiveness of LinProVSR's supervised disambiguation learning for ambiguous words. Specifically, we first statistically analyze the occurrence frequencies of ambiguous words, sort them by frequency in descending order to form the sorted set SA, then compute the averaged word error rate as follows:

$$\text{Avg}(AWER(m,n)) = \frac{1}{n-m+1} \sum_{i=m}^{n} AWER(i) \quad (17)$$

In Fig. 4, the x-axis label K denotes the first K ambiguous words in the sorted set SA; the left Y-axis illustrates the average error rate for these high-frequency words (Top-K), while the right Y-axis shows the reduction in average error rate achieved by LinProVSR relative to the baseline model. The findings reveal that as K decreases (indicating higher frequency of ambiguous words in training data), LinProVSR's correction advantage progressively strengthens, particularly achieving remarkable improvements in Top-5 high-frequency word recognition (with a 4.41% reduction on CMLR(a) and 5.32% reduction on LRS2(b)).

Conclusion

We address viseme-phoneme ambiguity in traditional lip reading by proposing LinProVSR, which consists of a linguistic prior-driven ambiguous sample construction module and a progressive contrastive disambiguation network. By imposing dual constraints via viseme-phoneme and text semantic contrastive losses across both feature and semantic spaces, the model extracts phoneme-related features while encoding and integrates linguistic insights into the feature learning and decoding processes of lip reading. Experiments on Chinese and English datasets demonstrate superior performance, thereby validating the model's effectiveness.

Acknowledgments

This research was partially supported by the following funding sources: the National Natural Science Foundation of China (Grant No. U24A20332, 62272143, 62476077), and the Seventh Special Support Plan for Innovation and Entrepreneurship in Anhui Province. The computation is completed on the HPC Platform of Hefei University of Technology.

References

Adam, K. D. B. J.; et al. 2014. A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 1412(6).

Afouras, T.; Chung, J. S.; Senior, A.; Vinyals, O.; and Zisserman, A. 2022. Deep Audio-Visual Speech Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12): 8717–8727.

Assael, Y. M.; Shillingford, B.; Whiteson, S.; and De Freitas, N. 2016. Lipnet: End-to-end sentence-level lipreading. *arXiv preprint arXiv:1611.01599*.

Bai, C.; Li, H.; Zhang, J.; Huang, L.; and Zhang, L. 2021. Unsupervised Adversarial Instance-Level Image Retrieval. *IEEE Transactions on Multimedia*, 23: 2199–2207.

Cheng, X.; Jin, T.; Huang, R.; Li, L.; Lin, W.; Wang, Z.; Wang, Y.; Liu, H.; Yin, A.; and Zhao, Z. 2023. Mixspeech: Cross-modality self-learning with audio-visual stream mixup for visual speech translation and recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15735–15745.

Chung, J. S.; Senior, A.; Vinyals, O.; and Zisserman, A. 2017a. Lip Reading Sentences in the Wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3444–3453. Honolulu, HI, USA: IEEE. ISBN 978-1-5386-0457-1.

Chung, J. S.; Senior, A.; Vinyals, O.; and Zisserman, A. 2017b. Lip Reading Sentences in the Wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3444–3453. Honolulu, HI, USA: IEEE. ISBN 978-1-5386-0457-1.

Cooke, M.; Barker, J.; Cunningham, S.; and Shao, X. 2006. An Audio-Visual Corpus for Speech Perception and Automatic Speech Recognition. *The Journal of the Acoustical Society of America*, 120(5): 2421–2424.

Do, J.; and Kim, M. 2025. SkateFormer: Skeletal-Temporal Transformer for Human Action Recognition. In Leonardis, A.; Ricci, E.; Roth, S.; Russakovsky, O.; Sattler, T.; and Varol, G., eds., *Computer Vision – ECCV 2024*, volume 15099, 401–420. Cham: Springer Nature Switzerland. ISBN 978-3-031-72939-3 978-3-031-72940-9.

Guo, W.; Sun, Y.; Xu, Y.; Qiao, Z.; Yang, Y.; and Xiong, H. 2024. SpGesture: Source-Free Domain-adaptive sEMG-based Gesture Recognition with Jaccard Attentive Spiking Neural Network. *Advances in Neural Information Processing Systems*, 37: 36717–36747.

Hao, B.; Zhou, D.; Li, X.; Zhang, X.; Xie, L.; Wu, J.; and Yin, E. 2025. LipGen: Viseme-Guided Lip Video Generation for Enhancing Visual Speech Recognition. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. Hyderabad, India: IEEE. ISBN 9798350368741.

Harte, N.; and Gillen, E. 2015. TCD-TIMIT: An Audio-Visual Corpus of Continuous Speech. *IEEE Transactions on Multimedia*, 17(5): 603–615.

Huang, Y.; Liang, X.; and Fang, C. 2021. CALLip: Lipreading Using Contrastive and Attribute Learning. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2492–2500. Virtual Event China: ACM. ISBN 978-1-4503-8651-7.

King, D. E. 2009. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10: 1755–1758.

Kit Khinn Teng, M.; Zhang, H.; and Saitoh, T. 2025. Phoneme-Level Visual Speech Recognition via Point-Visual Fusion and Language Model Reconstruction. *arXiv e-prints*, arXiv–2507.

Kumar, Y.; Sahrawat, D.; Maheshwari, S.; Mahata, D.; Stent, A.; Yin, Y.; Ratn Shah, R.; and Zimmermann, R. 2020. Harnessing GANs for Zero-Shot Learning of New Classes in Visual Speech Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(03): 2645–2652.

Laux, H.; Mededovic, E.; Hallawa, A.; Martin, L.; Peine, A.; and Schmeink, A. 2024. LITEVSR: Efficient Visual Speech Recognition by Learning from Speech Representations of Unlabeled Data. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 10391–10395. Seoul, Korea, Republic of: IEEE. ISBN 9798350344851.

Laux, H.; and Schmeink, A. 2024. Enhancing CTC-Based Visual Speech Recognition. *arXiv preprint arXiv:2409.07210*.

Li, Y.; Xue, F.; Guo, D.; Tang, S.; Li, P.; Li, S.; and Hong, R. 2025. CFLip: Generalizing Lipreading to Unseen Speakers by Learning Common Features. *IEEE Transactions on Computational Social Systems*, 1–16.

Li, Y.; Xue, F.; Wu, L.; Xie, Y.; and Li, S. 2024. Generalizing Sentence-Level Lipreading to Unseen Speakers: A Two-Stream End-to-End Approach. *Multimedia Systems*, 30(1): 42.

Ma, P.; Wang, Y.; Petridis, S.; Shen, J.; and Pantic, M. 2022. Training strategies for improved lip-reading. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8472–8476. IEEE.

Pan, X.; Chen, P.; Gong, Y.; Zhou, H.; Wang, X.; and Lin, Z. 2022. Leveraging unimodal self-supervised learning for multimodal audio-visual speech recognition. *arXiv preprint arXiv:2203.07996*.

Park, Y.-H.; Park, R.-H.; and Park, H.-M. 2025. Swinlip: An efficient visual speech encoder for lip reading using swin transformer. *Neurocomputing*, 130289.

Petridis, S.; Stafylakis, T.; Ma, P.; Tzimiropoulos, G.; and Pantic, M. 2018. Audio-Visual Speech Recognition with a Hybrid CTC/Attention Architecture. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, 513–520.

Prajwal, K. R.; Afouras, T.; and Zisserman, A. 2022. Sub-Word Level Lip Reading With Visual Attention. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5152–5162. New Orleans, LA, USA: IEEE. ISBN 978-1-66546-946-3.

Siddiqui, N.; Tirupattur, P.; and Shah, M. 2024. DVANet: Disentangling View and Action Features for Multi-View Action Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(5): 4873–4881.

Son, J. S.; and Zisserman, A. 2017. Lip Reading in Profile. In *Procedings of the British Machine Vision Conference 2017*, 155. London, UK: British Machine Vision Association. ISBN 978-1-901725-60-5.

Song, P.; Guo, D.; Yang, X.; Tang, S.; and Wang, M. 2024. Emotional Video Captioning With Vision-Based Emotion Interpretation Network. *IEEE Transactions on Image Processing*, 33: 1122–1135.

Wang, H.; Guo, P.; Wan, X.; Zhou, H.; and Xie, L. 2024. Enhancing Lip Reading with Multi-Scale Video and Multi-Encoder. In *2024 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 1–6. Niagara Falls, ON, Canada: IEEE. ISBN 9798350379815.

Wang, J.; Qian, X.; Zhang, M.; Tan, R. T.; and Li, H. 2023. Seeing what you said: Talking face generation guided by a lip reading expert. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14653–14662.

Weng, X.; and Kitani, K. 2019. Learning Spatio-Temporal Features with Two-Stream Deep 3D CNNs for Lipreading.

Wu, L.; Zhang, X.; Zhang, Y.; Zheng, C.; Liu, T.; Xie, L.; Yan, Y.; and Yin, E. 2024. Landmark-Guided Cross-Speaker Lip Reading with Mutual Information Regularization. *arXiv preprint arXiv:2403.16071*.

Xu, B.; Lu, C.; Guo, Y.; and Wang, J. 2020. Discriminative multi-modality speech recognition. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 14433–14442.

Xu, K.; Li, D.; Cassimatis, N.; and Wang, X. 2018. LCANet: End-to-End Lipreading with Cascaded Attention-CTC. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 548–555. Xi'an: IEEE. ISBN 978-1-5386-2335-0.

Xue, F.; Li, Y.; Liu, D.; Xie, Y.; Wu, L.; and Hong, R. 2023a. LipFormer: Learning to Lipread Unseen Speakers Based on Visual-Landmark Transformers. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9): 4507–4517.

Xue, F.; Yang, T.; Liu, K.; Hong, Z.; Cao, M.; Guo, D.; and Hong, R. 2023b. LCSNet: End-to-end Lipreading with Channel-aware Feature Selection. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 19(1s): 1–21.

Yang, S.; Gong, Z.; and Kang, J. 2023. An Improved End-to-End Audio-Visual Speech Recognition Model. In *INTERSPEECH 2023*, 3093–3097. ISCA.

Yeo, J. H.; Kim, M.; Choi, J.; Kim, D. H.; and Ro, Y. M. 2024. AKVSR: Audio Knowledge Empowered Visual Speech Recognition by Compressing Audio Knowledge of a Pretrained Model. *IEEE Transactions on Multimedia*, 26: 6462–6474.

Yu, J.; Zhang, S.-X.; Wu, J.; Ghorbani, S.; Wu, B.; Kang, S.; Liu, S.; Liu, X.; Meng, H.; and Yu, D. 2020. Audio-Visual Recognition of Overlapped Speech for the LRS2 Dataset. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6984–6988.

Zhang, F.; Zhu, Y.; Wang, X.; Chen, H.; Sun, X.; and Xu, L. 2024. Visual Hallucination Elevates Speech Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17): 19542–19550.

Zhang, H.; Zhu, B.; Cao, Y.; and Hao, Y. 2025. Hand1000: Generating realistic hands from text with only 1,000 images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 9, 9905–9913.

Zhang, X.; Gong, H.; Dai, X.; Yang, F.; Liu, N.; and Liu, M. 2019. Understanding Pictograph with Facial Features: End-to-End Sentence-Level Lip Reading of Chinese. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01): 9211–9218.

Zhao, Y.; Xu, R.; and Song, M. 2019. A Cascade Sequence-to-Sequence Model for Chinese Mandarin Lip Reading. In *Proceedings of the ACM Multimedia Asia*, 1–6. Beijing China: ACM. ISBN 978-1-4503-6841-4.

Zhao, Y.; Xu, R.; Wang, X.; Hou, P.; Tang, H.; and Song, M. 2020. Hearing Lips: Improving Lip Reading by Distilling Speech Recognizers. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04): 6917–6924.

Zhou, Z.; Zhao, G.; Hong, X.; and Pietikäinen, M. 2014. A Review of Recent Advances in Visual Speech Decoding. *Image and Vision Computing*, 32(9): 590–605.