Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Multi-talker audio–visual speech recognition towards diverse scenarios*&

Yuxiao LIN, Tao JIN, Xize CHENG, Zhou ZHAO, Fei WU‡

College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China

E-mail: yuxiaolinling@zju.edu.cn; jint_zju@zju.edu.cn; chengxize@zju.edu.cn; zhaozhou@zju.edu.cn; wufei@zju.edu.cn

Received June 13, 2025; Revision accepted Nov. 2, 2025; Crosschecked Nov. 17, 2025; Published online Dec. 8, 2025

Abstract: Recently, audio–visual speech recognition (AVSR) has attracted increasing attention. However, most existing works simplify the complex challenges in real-world applications and only focus on scenarios with two speakers and perfectly aligned audio-video clips. In this work, we study the effect of speaker number and modal misalignment in the AVSR task, and propose an end-to-end AVSR framework under a more realistic condition. Specifically, we propose a speaker-number-aware mixture-of-experts (SA-MoE) mechanism to explicitly model the characteristic difference in scenarios with different speaker numbers, and a cross-modal realignment (CMR) module for robust handling of asynchronous inputs. We also use the underlying difficulty difference and introduce a new training strategy named challenge-based curriculum learning (CBCL), which forces the model to focus on difficult, challenging data instead of simple data to improve efficiency.

Key words: Speech recognition and synthesis; Multi-modal recognition; Curriculum learning; Multi-talker speech recognition

https://doi.org/10.1631/FITEE.2500411 **CLC number:** TP18

1 Introduction

Speech recognition in multi-talker scenarios remains one of the most challenging tasks in the speech processing community. Although modern systems have achieved human-level performance on clean speech benchmarks and demonstrated remarkable robustness against background noise (Xiong et al., 2016; Nguyen et al., 2021), their performance degrades significantly in the presence of overlapping speech. This limitation poses substantial barriers to practical applications such as meeting transcription systems, where spontaneous multi-speaker interactions are common. In response, recent research has increasingly focused on multi-talker speech separation (Hershey et al., 2016) and recognition (Gulati et al., 2020; Guo et al., 2021), aiming to extend the applicability of speech technologies to broader real-world settings.

One promising research direction leverages the visual modality through audio–visual speech recognition (AVSR) (Gao and Grauman, 2021; Cheng XZ et al., 2023; Yang XD et al., 2024). Fueled by the availability of large-scale audio–visual datasets (Nagrani et al., 2017; Afouras et al., 2018a, 2018b), AVSR extends traditional automatic speech recognition (ASR) into a multi-modal framework by leveraging visual cues from lip movements. This approach offers two key advantages: It enables artificial intelligence (AI) systems to perceive speech more similarly to humans, and it naturally resolves the permutation ambiguity inherent to multi-talker ASR (Yu et al., 2017) by visually specifying the target speaker. Although many existing AVSR studies focus on

‡ Corresponding author

* Project supported by the National Natural Science Foundation of China (No. 62572423)

& A preliminary version was presented at the IEEE International Conference on Acoustics, Speech and Signal Processing, April 6–11, 2025, India

ORCID: Yuxiao LIN, https://orcid.org/0000-0002-3954-5927; Fei WU, https://orcid.org/0000-0003-2139-8807

improving performance in clean audio scenarios (Ma et al., 2021; Haliassos et al., 2023), recent works have begun addressing speech noise challenges, primarily focusing on single intervening speaker cases (Shi et al., 2022b; Ma et al., 2023).

However, real-world applications present more complex challenges that current research has yet to fully address. First, the number of overlapping speakers involved in a recording is often unknown. Although some audio–visual speech separation systems train on data with varying speaker counts (Cheng HY et al., 2023), they typically treat this as simple data augmentation without explicitly studying how different speaker numbers affect recognition performance. We argue that the impact of speaker counts should be systematically considered, as optimal modality utilization varies with scenario. In simpler scenarios with no intervening speakers, audio alone may suffice, whereas visual information becomes increasingly crucial as speaker overlap grows. To address this dynamic requirement, we propose a speaker-number-aware mixture-of-experts (SA-MoE) mechanism. This architecture explicitly assigns different scenarios to different experts while using speaker counting as an auxiliary task for automatic expert assignment. By incorporating low-rank decomposition (Hu EJ et al., 2022), SA-MoE achieves this adaptive capability with minimal additional parameters.

Second, practical AVSR systems must address the modality alignment challenge. Given the scarcity of annotated data, many works use unlabeled audio–visual datasets to pre-train multi-modal encoders to boost downstream AVSR tasks (Shi et al., 2022a; Haliassos et al., 2023; Zhu et al., 2024). Although most pre-training models include scenarios with missing modalities during training, these approaches typically assume perfect temporal alignment between modalities when both are available, overlooking real-world scenarios where hardware limitations, asynchronous recording setups, or environmental factors induce modality misalignment. This work substantially extends our prior research (Lin et al., 2025), which was operated under the simplified assumption of perfectly aligned modalities and did not address real-world modality asynchrony. To address this, we propose a cross-modal realignment (CMR) module that integrates with pre-trained models to automatically correct input sequence misalignment.

Furthermore, these challenges naturally divide scenarios into groups of varying difficulty levels, making them well-suited for curriculum learning to further enhance model performance. By progressively training the model—starting from easier data with fewer speakers and perfect alignment before advancing to more challenging cases—we enable more effective learning. However, we identify a key limitation in standard curriculum learning: Easier data often dominate the training process at the expense of harder examples. To address this, we propose challenge-based curriculum learning (CBCL), a novel training strategy where the proportion of easy data decreases exponentially as training progresses. This approach reduces excessive focus on simpler cases while ensuring sufficient attention to harder scenarios, ultimately improving overall performance.

Our work makes three primary contributions:

1. The SA-MoE architecture for adaptive modality utilization;

2. The CMR module for robust handling of asynchronous inputs;

3. The CBCL strategy for optimized training on multi-speaker scenarios.

Together, these innovations advance AVSR's applicability to complex real-world scenarios with varying speaker counts and imperfect recording conditions.

2 Related works

2.1 AVSR

Research in AVSR has spanned decades, but deep learning-based approaches have gained significant attention only recently, enabled by the availability of large-scale annotated audio–visual datasets (Chung et al., 2017; Afouras et al., 2018a, 2018b; Yang S et al., 2019). Given the scarcity of transcribed data, many studies leverage unlabeled datasets like VoxCeleb (Nagrani et al., 2017) to pre-train multi-modal encoders for enhanced representation learning. Notable examples include: AV-HuBERT (Shi et al., 2022a, 2022b), which introduces a self-supervised framework for audio–visual speech; VatLM (Zhu et al., 2024), using both paired and unpaired multi-modal data; Auto-AVSR (Ma et al., 2023), employing pre-trained ASR models to generate transcriptions for unlabeled audio–visual

corpora.

Most AVSR research focused on designing architectures that effectively integrate visual and auditory modalities (Afouras et al., 2018a; Petridis et al., 2018; Makino et al., 2019; Ma et al., 2021). These efforts primarily target performance improvements in clean audio scenarios or cases with single intervening speakers. Although some audio–visual speech separation systems train on data with varying speaker counts (Cheng HY et al., 2023), they typically treat speaker diversity merely as data augmentation. In contrast, our approach explicitly differentiates processing strategies based on speaker count.

For scenarios with temporally misaligned data, some models process modalities as independent sequences (Rahimi et al., 2022; Li et al., 2024), using audio–visual cross-attention to establish intermodal relationships while circumventing alignment issues. However, this sacrifices explicit modeling of audio–visual correlations. Other works design specialized encoders for misaligned inputs (Hu YC et al., 2023), but their structural modifications alter feature spaces, making them incompatible with existing pre-trained models. Our method uniquely preserves inherent audio–visual correlations while maintaining compatibility with pre-trained knowledge.

2.2 Curriculum learning

Human education is highly organized, starting from easy, intuitive concepts and increasing the difficulty as the learning progresses. Curriculum learning, first proposed by Bengio et al. (2009), is the idea of arranging training data in a meaningful order, similar to human learning.

Due to the lack of a standard criterion, difficulty measurement is the main challenge of applying curriculum learning to many tasks. As a result, many automatic measurement methods have been proposed, including the usage of self-paced learning (Kumar MP et al., 2010), transfer learning (Weinshall et al., 2018; Hacohen and Weinshall, 2019), and reinforcement learning (Graves et al., 2017; Kumar G et al., 2019). The other main focus of curriculum learning is to schedule the training procedure effectively given the difficulty criterion (Spitkovsky et al., 2010; Platanios et al., 2019; Penha and Hauff, 2020).

3 Methods

3.1 Task overview

Given a speech mixture containing an unknown number of simultaneous speakers and the video stream of the target speaker that might be temporally misaligned, our goal is to acquire the transcription of the target speaker. The models are trained on synthesized audio mixtures generated from single-speaker datasets due to the scarcity of speech mixture data with video and transcription. As a result, during the training phase, the speaker number and time shift in each data point are known, making it possible to form an auxiliary task and design training strategies based on speaker numbers to improve model performance in scenarios with different speaker numbers.

Formally, each data point consists of a tuple $(\boldsymbol{V}, \boldsymbol{A}, \boldsymbol{w}, n, f)$, where $\boldsymbol{V} \in \mathbb{R}^{T\times w\times h}$ is the target video stream of size $w \times h$ and length T, $\boldsymbol{A} \in \mathbb{R}^{T\times c}$ is the acoustic feature extracted from the mixture $\boldsymbol{a} = \sum_{i=1}^{n} \boldsymbol{a}_i$ generated using target speech $\boldsymbol{a}_1$ and $n-1$ intervene speeches $\boldsymbol{a}_2$ to $\boldsymbol{a}_n$, and $\boldsymbol{w} = (w_1, w_2, ..., w_s)$ is the token sequence of the ground-truth transcription. n is the number of speeches used in the mixture, and f is an integer representing the frame shift between audio and video modality. A positive f indicates that audio is played earlier than video, whereas a negative f means later. The main task of the model is to predict the transcription $\hat{\boldsymbol{w}} = \text{Model}(\boldsymbol{V}, \boldsymbol{A})$ given a target video and mixed audio input.

3.2 Backbone structure

We adopt the same Transformer-based encoder-decoder structure as AV-HuBERT, which is used in various pre-training works (Shi et al., 2022b; Haliassos et al., 2023; Zhu et al., 2024) as the backbone model. It allows us to fully use the knowledge in the pre-trained model. As shown in Fig. 1, audio and video features first pass through their respective frame-wise lightweight encoder to produce intermediate features, which are then concatenated and fed into a fusion layer to obtain multi-modal input features $\boldsymbol{H}^{(0)} \in \mathbb{R}^{T\times C}$ for the SA-MoE encoder:

$$\boldsymbol{H}^{(0)} = \text{Fusion}([\text{ResNet}(\boldsymbol{V}); \text{FFN}(\boldsymbol{A})]), \quad (1)$$

where C is the embedding dimension, and $[\cdot;\cdot]$ indicates concatenation over feature dimension.

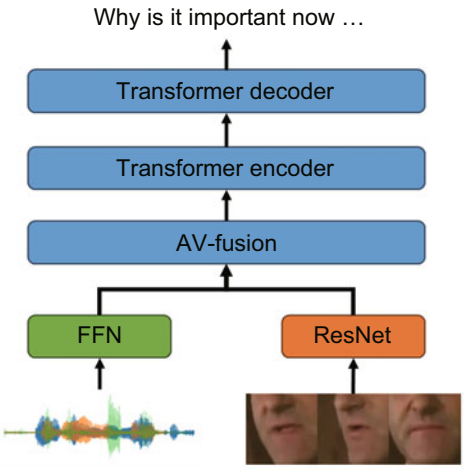

Fig. 1 Backbone structure. FFN: feed forward network

The K-layer encoder takes it as input and generates a latent feature sequence:

$$\boldsymbol{H}^{(K)} = \mathrm{Enc}(\boldsymbol{H}^{(0)}). \tag{2}$$

The Transformer decoder then auto-regressively decodes the feature sequence into target probabilities and computes the prediction loss:

$$L_{\mathrm{pred}} = -\sum_{t=1}^{s} \ln p(w_t | w_{1:t-1}, \boldsymbol{H}^{(K)}). \tag{3}$$

All modules except the decoder are initialized with parameters pre-trained on large-scale, unlabeled datasets.

3.3 SA-MoE encoder

Although the tasks of transcribing texts from an audio mixture with different speaker numbers are closely related, there exist distinct characteristics in each scenario. For example, in a 2-mix scenario, models might be able to transcribe from audio directly and only need the video information to locate the target speaker, whereas in a noisy 4-mix scenario, it becomes necessary for the model to perform lip-reading to get high-quality results. This raises concerns that a single unified model might not be flexible enough to generalize to all scenarios. To overcome the problem, we adopt the mixture-of-experts (MoE) structure that has been used in many multi-lingual speech recognition models (Kwon and Chung, 2023; Wang et al., 2023), and propose an SA-MoE encoder to explicitly assign different model components to different scenarios.

The key components of the SA-MoE encoder include a router and multiple MoE layers. As shown in Fig. 2, starting from a standard Transformer encoder with K blocks, we feed the output of the k^{th} block to the router to determine which experts to activate, and replace all linear layers in the following $K-k$ Transformer blocks with MoE layers that take the router's output as extra information.

The router takes the speaker count problem as a classification task, and generates frame-wise probabilities for speaker number prediction with the output of the encoder's k^{th} Transformer layer $\boldsymbol{H}^{(k)}$. A negative log-likelihood (NLL) loss is computed and added to the total model loss for optimization.

Linear layers in the remaining $K-k$ Transformer blocks are replaced by an MoE structure. The structure consists of a shared linear layer and N expert layers, as shown in Fig. 3. All the experts have the same input and output dimensions as the original linear layer. The output of the structure is the sum of the output of the shared layer and the weighted sum of the expert outputs:

$$\boldsymbol{y} = \boldsymbol{W}_s \boldsymbol{x} + \sum_{i=1}^{N} s_i \boldsymbol{W}_i \boldsymbol{x}, \tag{4}$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ are the input and output of the layer of size C_{in} and C_{out}, respectively, $\boldsymbol{W}_s, \boldsymbol{W}_1, \boldsymbol{W}_2, \ldots, \boldsymbol{W}_N \in \mathbb{R}^{C_{\mathrm{out}} \times C_{\mathrm{in}}}$ are the weight

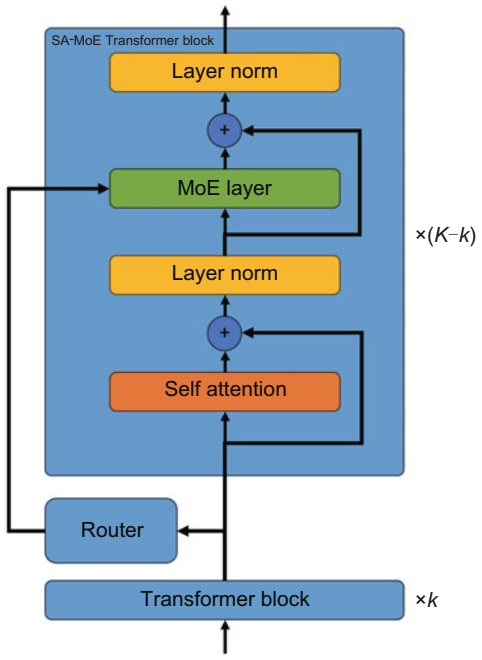

Fig. 2 Structure of SA-MoE

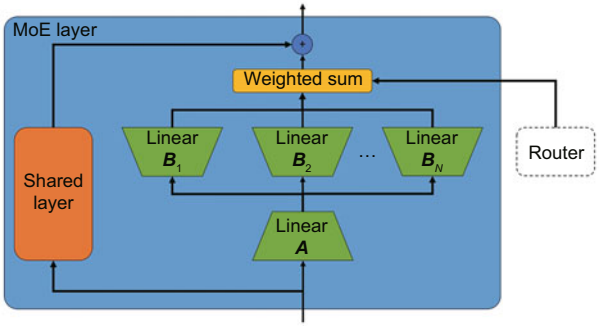

Fig. 3 MoE layer

matrices of the shared layer and experts, respectively, and $\boldsymbol{s} = (s_1, s_2, ..., s_N)$ is the output of the router. As linear layers are applied in a frame-wise manner, we omit the time dimension here. Guided by the router outputs, each layer can efficiently learn scenario-specific knowledge, and the shared layer contains the common knowledge for the task of multi-speaker speech recognition. It is worth noticing that, unlike approaches in multi-lingual speech recognition where sparse MoE methods are used, our router performs a weighted sum, which allows multiple routes to activate at a time. In multi-lingual tasks, the boundary between different languages is clear, whereas in our task, the scenario shift is continuous to some extent. For example, a 3-speaker mixture with fewer overlaps can be closer to the characteristics of a 2-speaker scenario. Using a soft router allows the model to recognize in a more flexible way, which helps improve the generalizability.

However, replacing every linear layer with $N+1$ layers of the same size increases the model size significantly, and activating all routes together introduces extra computation cost compared to sparse approaches. Because each route only corresponds to a portion of the training data, the model can be too large for effective training. To build a lightweight route layer, we adopt the concept in LoRA (Hu EJ et al., 2022) and replace the weight matrix of each expert with a low-rank decomposition $\boldsymbol{W}_i = \boldsymbol{B}_i\boldsymbol{A}_i$, where $\boldsymbol{A}_i \in \mathbb{R}^{r\times C_{\text{in}}}$, $\boldsymbol{B}_i \in \mathbb{R}^{C_{\text{out}}\times r}$, and the rank $r \ll \min(C_{\text{out}}, C_{\text{in}})$. Furthermore, we share the first matrix among all experts, $\boldsymbol{A}_1 = \boldsymbol{A}_2 = \cdots = \boldsymbol{A}_N = \boldsymbol{A}$. Finally, the modified linear layer yields

$$\boldsymbol{y} = \boldsymbol{W}_s\boldsymbol{x} + \sum_{i=1}^{N} s_i\boldsymbol{B}_i\boldsymbol{A}\boldsymbol{x}. \tag{5}$$

As linear layers are applied in a frame-wise manner, we omit the time dimension here. Using low-rank decomposition, the extra parameter introduced only contributes to $\left(\frac{1}{C_{\text{out}}} + \frac{N}{C_{\text{in}}}\right) r$ times the original parameters. With a small r, the cost is negligible.

This design is compatible with any model structure and allows us to use the feature extraction ability of existing pre-trained models. By initializing $\boldsymbol{W}_s$ with the pre-trained parameters of the original linear layer and all $\boldsymbol{B}_i$ with zero, the modified model retains the same behavior as the original pre-trained model at the beginning of training.

3.4 Cross-modal realignment

In audio–visual pre-training models (Shi et al., 2022a; Haliassos et al., 2023; Zhu et al., 2024), input features of different modalities are usually fused through frame-wise addition or concatenation. When a modality is missing, the corresponding inputs are set to 0. Consequently, the model develops a unified feature extraction capability that handles both single-modality and multi-modality inputs. However, in scenarios with misaligned data, features belonging to different frames are forcibly fused, raising the concern that it may lead to a degradation of the quality of the extracted features, which in turn reduces the overall performance.

Our work introduces a CMR mechanism that can reconstruct aligned input sequences to recover the explicit correlation and fully use pre-trained models throughout the process.

Fig. 4 shows the main workflow of the mechanism. Instead of directly concatenating the per-frame input audio features $\boldsymbol{X}_{\text{a}}$ and video features $\boldsymbol{X}_{\text{v}}$, we use a CMR module to reconstruct an aligned video feature sequence before extracting multi-modal features for speech recognition.

We first use the encoder to extract multi-modal features $\boldsymbol{H}_{\text{av}}$ and single-modal video features $\boldsymbol{H}_{\text{v}}$.

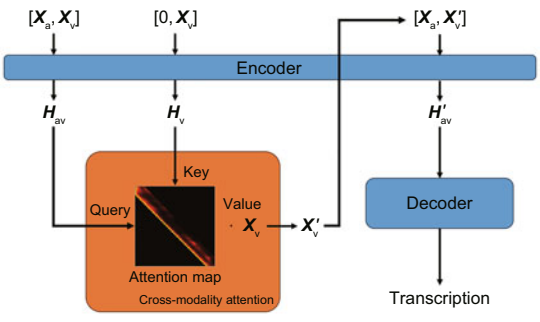

Fig. 4 Workflow of CMR

After that, we feed them into a cross-attention layer with $\boldsymbol{H}_{\mathrm{av}}$ as query and $\boldsymbol{H}_{\mathrm{v}}$ as key to obtain attention weights $\boldsymbol{W}_{\mathrm{attn}} \in \mathbb{R}^{(t\times t)}$:

$$\boldsymbol{W}_{\mathrm{attn}_i} = \mathrm{softmax}\left(\frac{(\boldsymbol{W}_{\mathrm{q}_i}\boldsymbol{H}_{\mathrm{av}})(\boldsymbol{W}_{\mathrm{k}_i}\boldsymbol{H}_{\mathrm{v}})^{\mathrm{T}}}{\sqrt{d_{\mathrm{k}}}} + \boldsymbol{M}\right),$$
$$\boldsymbol{W}_{\mathrm{attn}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{W}_{\mathrm{attn}_i}, \tag{6}$$

where $\boldsymbol{W}_{\mathrm{q}_i}$ and $\boldsymbol{W}_{\mathrm{k}_i}$ are the weights of the projection matrix for query and key in the i^{th} head, respectively, d_{k} is the dimension of key, and $\boldsymbol{M}$ is an attention mask that will be discussed later. We then multiply the weights by the original video feature to obtain a new video feature:

$$\boldsymbol{X}'_{\mathrm{v}} = \boldsymbol{W}_{\mathrm{attn}}\boldsymbol{X}_{\mathrm{v}}. \tag{7}$$

In contrast to the conventional multi-head attention mechanism, we omit the value projection on $\boldsymbol{X}_{\mathrm{v}}$ prior to attention computation. As a result, the output $\boldsymbol{X}'_{\mathrm{v}}$ can be viewed as a linear combination of the original visual input features. This ensures that new features remain in the same feature space as the original, and knowledge learned in the pre-trained encoder can still be applied to the new features.

We use the shared pre-trained encoder throughout the procedure. In this way, we use the model's ability to extract information from single- and multi-modal input. It also serves as a regularization, which results in better generalization in aligned and misaligned scenarios.

The reason we use $\boldsymbol{H}_{\mathrm{av}}$ instead of $\boldsymbol{H}_{\mathrm{a}}$ as the query is that, unlike AVSR with a single speaker, audio features in multi-talker scenarios are severely corrupted, and extracting features from audio only can be hard or even impossible. Using multi-modal features as the query ensures the ability to extract meaningful $\boldsymbol{X}'_{\mathrm{v}}$ even under extremely noisy scenarios.

Learning the audio–visual alignment implicitly is hard, so we add an auxiliary task to aid model training. The attention weight matrix $\boldsymbol{W}_{\mathrm{attn}}$ can be viewed as the results of a classification task choosing the matching video frame given an audio frame. For data point with frame shift k, we generate attention label sequence $\boldsymbol{a} = (a_1, a_2, \ldots, a_T)$:

$$a_i = \begin{cases} i - f, & 0 \le i - f < T, \\ -1, & \text{otherwise.} \end{cases} \tag{8}$$

Here, -1 indicates that the frame is at the beginning or end of the input sequence where no corresponding video frame is available, so we ignore these frames for loss computing. We then compute a cross-entropy loss and add it as an auxiliary loss:

$$L_{\mathrm{attn}} = \mathrm{CrossEntropy}(\boldsymbol{W}_{\mathrm{attn}}, \boldsymbol{a}). \tag{9}$$

Because attention is a global operation involving the whole sequence, an audio frame may attend to video frames far from it on the time axis, but with similar content, which is not desirable in our frame shift prediction task. We add a diagonal mask $\boldsymbol{M}$ to the attention operation to solve this. $\boldsymbol{M}$ is a matrix of size $t \times t$ where

$$M_{i,j} = \begin{cases} 0, & |i - j| \le F, \\ -\inf, & \text{otherwise.} \end{cases} \tag{10}$$

Here, F is the maximum expected frame shift in the experimental setting. With the mask, every audio feature will only attend to video features at most F frames away from it.

When audio falls behind video, audio corresponding to video frames at the end of a clip may be outside the clip. As a result, there will be no query in the audio sequence to retrieve them. Video frames at the beginning of the clip also tend to be discarded when a video falls behind audio for the same reason. However, these video frames can still carry important visual clues, especially when the first or last word in the audio segment is incomplete. To allow the model to keep the information, we pad F frames at the beginning and the end of input audio and video features. In this way, the dummy audio frames can help collect the information from these frames.

Fig. 5 gives an example of computing cross-modal attention. Note that while the padded input frames contain no information, they can still receive contextual information after going through the encoder, which helps identify the corresponding frame in the visual modality.

3.5 Challenge-based curriculum learning

Curriculum learning is a type of learning strategy, in which learning starts with mini-batches sampled only from easy data and increases the difficulty gradually. The strategy mimics the learning pattern of humans, and can help the model achieve a better local optimum. Curriculum learning involves two

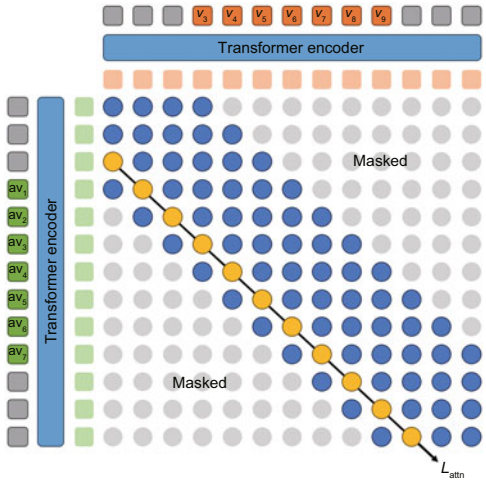

Fig. 5 An example of computing the cross-modal attention with $F = 3$ and $f = -2$. Gray squares denote dummy frames padded to the input sequence, and gray circles denote masked attention. The attention loss is computed along the yellow circles. References to color refer to the online version of this figure

basic components: a scoring function to measure the difficulty of each element in dataset and sort them in ascending order, and a monotonically increasing pacing function that determines a subset of the full dataset where each is sampled. Although defining an appropriate scoring function is the main challenge for most curriculum learning tasks, we naturally have the prior knowledge that the difficulty scales with speaker number and frame shift, which makes it especially suitable for our task.

When establishing a human curriculum, the teacher will focus more on hard materials after students master the easy ones, because hard materials also contain knowledge learned from easy materials. Although the curriculum will include reviewing learned materials from time to time, the main focus is always on hard and unseen materials. However, this is not the case in the standard curriculum learning framework, where mini-batches are uniformly sampled from the subset determined by the pacing function. At every stage, the easiest data have the same weight as the current hardest data, and the hardest data in the whole dataset will only be selected during the last training stages. Instead of focusing on hard samples, the model input is inevitably biased towards easy data during the whole training procedure. This can lead to suboptimal performance on hard samples.

To address this issue, we propose a challenge-based curriculum learning (CBCL) strategy, which always focuses on the hardest data in the current subset. We start with training the model in one scenario and add harder scenarios into the training data stage by stage. At each stage, there is a chance of 50% that a data point is sampled from the current highest difficulty, while the other 50% is sampled from the distribution of the previous stage. In this way, the weight of easy scenarios decreases exponentially, and the model can always focus on hard, challenging data.

The first row in Fig. 6 illustrates the ratio of different data in different training strategies. Notice that in the standard curriculum learning procedure, the hardest data only contribute to 1/4 of the data at the last training stage, which is only a small portion considering the whole training process. With our strategy, the share of the hardest data is doubled. Although the hardest data still contribute to a relatively small portion throughout the whole training process, the situation can change if we take into consideration the time dimension. For both human and neural networks, newly learned information usually has much more impact than material learned earlier. Studies in neuroscience usually model the effect as an exponential forgetting curve (Ebbinghaus, 1885; Murre and Dros, 2015). If we also weight the influence of training data by an exponential function as shown in the second row in Fig. 6, the significance of hard data increases substantially at the final stage, as it accounts for the majority of the processing.

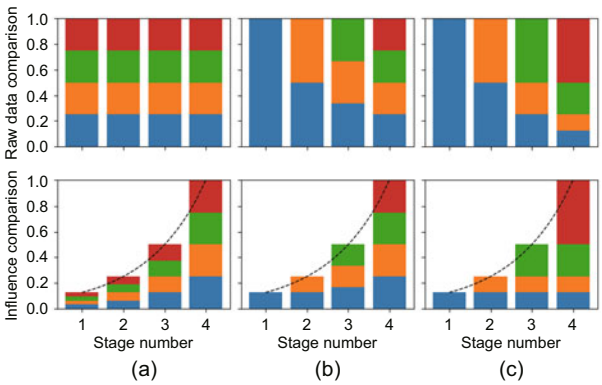

Fig. 6 Comparison of data ratio using different training strategies: (a) mix training; (b) standard CL; (c) CBCL. Different colors indicate data added at different stages of the curriculum. Figures on the top row show the ratio of raw data at each stage, and the bottom row illustrates the effect of the data on the model, assuming the effect decreases exponentially over time

4 Experiments

4.1 Data and experimental setup

We evaluate model performances on LRS3 (Afouras et al., 2018b), the largest publicly available audio–visual speech dataset with transcriptions. The dataset is collected from TED and TEDx videos and contains 433 hours of data from more than 5000 speakers.

To generate a multi-talker mixture, we first sample intervention speeches and crop them to the length of the target speech. Target speech and intervention speeches are then normalized to the same energy level and added together. During training, we randomly select 1 to 4 speakers, whereas during evaluation, we further include 5-speaker mixtures to evaluate the generalization of our model. To simulate the frame shift, we set F to 5 and sample a time shift value f between -5 and 5 frames (-0.2 s to 0.2 s). If $f > 0$, we remove the first f frames of audio features and the last f frames of video features; if $f < 0$, we remove the last f frames of audio features and the first f frames of video features.

We adopt the same encoder structure as AV-HuBERT base (Shi et al., 2022a, 2022b) as our backbone, and initialize the model with its pre-trained parameters. It uses a 26-dimensional log filter-bank energy feature at a stride of 10 ms as an audio feature and stacks four neighboring frames, so the audio feature is at the same sample rate as the 25 Hz video input. The model uses a modified ResNet-18 for extracting video features and a linear layer for audio features before feature fusion. The encoder contains 12 Transformer blocks, and the decoder contains six. For our SA-MoE encoder structure, the router takes the output feature of the second Transformer block as input, and all linear layers in the rest blocks are replaced with MoE layers of rank one. The backbone model contains 1.606×10^8 parameters, and models with SA-MoE + CMR structure contain 1.637×10^8.

All experiments are conducted using a single NVIDIA GeForce RTX 3090 GPU for 180 000 steps with update frequency set to 2. During the first 20 000 steps, we freeze the encoder with pre-trained parameters and only train the decoder. For all curriculum learning experiments, we split the training into four stages with 90 000 steps during the first stage and 30 000 during each of the other stages. Although data with only one speaker are technically the easiest case, it is known that a model trained on clean data can hardly generalize to multi-speaker scenarios. The 1-speaker scenario should be considered as a degraded special case of the task instead of a base scenario. Instead, the 2-speaker scenario is the fundamental case in our task. To allow the model to learn fundamental knowledge of the AVSR task at the beginning of training, we add data to the training set in the order of $(2, 1, 3, 4)$-speaker instead of $(1, 2, 3, 4)$.

4.2 Scenarios with different speaker numbers

We showcase the challenge posed by arbitrary speaker numbers by training several AV-HuBERT base models using datasets with a fixed speaker number, and evaluating them on datasets with different speaker numbers. As shown in Table 1, the model trained on 1-speaker data performs poorly on all multi-speaker scenarios. This is because the model lacks the knowledge of recognizing the target from the audio mixture. For other training settings, different models have similar test performances, which proves the existence of a relationship among different scenarios. Models trained on data with the same speaker number as the test setting tend to have the best performance, and perform worse as the difference in speaker number becomes larger. This proves the existence of distinct characteristics in each scenario. However, models trained on 4-speaker and 5-speaker data are outperformed by other models in their respective test scenarios. This illustrates the difficulty of training directly using data with a large number of speakers and indicates the need for curriculum learning, which starts training from an easier scenario.

In Table 2 we evaluate the effect of SA-MoE structure and CBCL training strategy using the

Table 1 WER of backbone models trained with fixed speaker number

N_{tr}	Performance				
	N_{te}=1	2	3	4	5
1	**2.13**	6.42	16.54	29.74	38.17
2	2.34	**4.01**	9.24	16.85	22.91
3	2.51	4.45	**8.19**	**13.11**	17.41
4	2.75	4.98	8.43	13.27	**17.18**
5	2.82	5.70	9.86	14.18	17.82

N_{tr} means the training speaker number. N_{te} means the testing speaker number. Best results are in bold. The gray diagonal pattern indicates that the training speaker numbers and testing speaker numbers are consistent

Table 2 WER combining different structures and training strategies

Training strategy	SA-MoE	Performance					Overall
		N_{sp}=1	2	3	4	5	
Mix	×	2.15±0.04	4.36±0.03	**8.03±0.13**	13.54±0.07	17.81±0.36	9.18±0.18
	✓	**2.08±0.12**	4.37±0.04	8.46±0.06	**12.81±0.07**	17.48±0.25	9.04±0.13
Std. CL	×	2.19±0.06	4.44±0.07	8.22±0.01	14.04±0.09	17.93±0.15	9.36±0.09
	✓	2.19±0.12	4.74±0.06	8.30±0.05	13.36±0.04	17.70±0.20	9.26±0.11
CBCL	×	2.23±0.04	4.83±0.08	8.22±0.07	13.20±0.05	17.46±0.32	9.19±0.15
	✓	2.15±0.07	**4.27±0.09**	8.18±0.08	12.98±0.07	**16.68±0.40**	**8.85±0.18**

N_{sp} means the speaker number. Best results are in bold. The underline means the second-best result

AV-HuBERT base backbone by training modes with or without the SA-MoE structure using different strategies.

Without SA-MoE, the model trained with the standard curriculum learning (Std. CL) strategy has similar performance on tests with 1 and 2 speakers compared to the model trained directly on mixed data (Mix), but the results become much worse as the speaker number grows, leading to a worse overall result. This shows the drawback of the strategy that the model spends more time on easy data that can be mastered with less effort and does not have enough time to learn hard data that requires more effort. With our CBCL training strategy, the model pays more attention to unseen harder scenarios and spends less time reviewing learned scenarios, which leads to a performance increase on tests with more speakers. However, the performance on easy scenarios drops due to the biased training set.

The proposed SA-MoE structure improves the performance under all training strategies, and the improvement is more significant when coupled with curriculum learning-based strategies. The phenomenon suggests that the structure performs better with curriculum learning than with direct training on mixed data. When directly trained on mixed data, the router may not be able to predict the speaker number correctly and activate the respective route at the beginning of the training. As a result, all routes may still learn shared knowledge together instead of scenario-specific ones during an early stage. In curriculum learning cases, the training process starts with data from a single scenario, which allows the router to learn to activate the corresponding expert quickly. When data of harder scenarios are added into the training set, the model tends to assign them to the nearest route before learning to correctly classify them as a new class. During the process, experts who are close to the new scenario can learn from it to improve generalizability, while the more distant experts are protected by the router and have a better chance to retain their scenario-specific knowledge. As a result, with curriculum learning, the multi-route model better retains the knowledge learned in 1-speaker and 2-speaker scenarios, and with the CBCL strategy, where it further improves performance with a larger number of speakers, it achieves the best overall performance. Moreover, despite being trained only on a dataset with 1 to 4 speakers and not having a route for 5-speaker data, CBCL with SA-MoE achieves much better performance compared to single-scenario experiments. This indicates the model's capability of generalizing learned scenario-specific knowledge to unseen scenarios and will not be hindered by the lack of the corresponding route.

Table 3 shows the model performance with expert rank 16 and various numbers of SA-MoE layers. On top of the word error rate (WER), we also report the accuracy of the speaker number prediction task for 1 to 4 speakers. Although the speaker number prediction becomes less accurate as the number of SA-MoE layers grows, recognition performance improves. Intermediate features closer to inputs contain more acoustic information, while those closer to outputs contain more linguistic components. As a result, to separate the target speech, experts should start intervening at an early stage of feature extraction. However, predicting speaker numbers without any context information will significantly reduce the accuracy (from 81.50% to 61.29%) and will affect the overall performance.

Table 4 shows the effect of rank on the model performance. As the result illustrates, lightweight experts with the rank set to 1 outperform experts with more parameters. The phenomenon suggests

Table 3 Model performance with different numbers of SA-MoE layers

Number of SA-MoE layers	Router accuracy (%)	Performance					Overall
		N_{sp}=1	2	3	4	5	
2	**91.33**	2.17	4.63	8.16	13.51	18.01	9.30
4	91.25	2.35	4.53	8.36	13.48	17.46	9.24
6	90.46	2.03	4.50	8.33	13.70	17.57	9.23
8	88.45	2.24	4.56	**7.84**	13.58	17.50	9.14
10	81.50	2.04	4.44	8.17	**13.34**	**17.30**	**9.06**
12	61.29	**2.02**	**4.33**	8.23	13.55	17.46	9.12

N_{sp} means the speaker number. Best results are in bold. The underline means the second-best result

Table 4 Model performance with different expert ranks

Expert rank	Performance					Overall
	N_{sp} =1	2	3	4	5	
1	2.24	**4.19**	8.13	**12.98**	**16.26**	**8.76**
4	2.17	4.26	**8.03**	13.11	16.97	8.91
16	**2.04**	4.44	8.17	13.34	17.30	9.06
64	2.22	4.79	8.43	13.35	17.81	9.32
256	2.67	6.45	10.18	16.26	19.64	11.04

N_{sp} means the speaker number. Best results are in bold. The underline means the second-best result

that scenarios with different players are closely related, and the differences can be learned with a simple model structure. Lightweight experts also enforce the model to use shared layers to learn common features across different scenarios, which improves the generalizability. Experts with more parameters are harder to optimize because the router only assigns part of the training data to each expert.

To illustrate the different knowledge learned by different experts, we manually set the router's outputs to one-hot labels during evaluation and report the results in Table 5. As the result shows, every expert performs the best when testing speaker number matches, and the results are worse in distant scenarios. The performance of the full model is close to each expert on their respective scenario, showing its ability to activate the correct expert based on inputs.

Existing works mainly study the performance of their AVSR models on 2-speaker scenario under different signal-to-noise ratio (SNR) levels. Using experiments, we illustrate that rescaling to different SNR levels is inadequate for evaluating the model. In our main experiments, speeches are normalized to the same energy level when generating mixtures. For mixtures with 2 to 5 speakers, their SNR levels are 0, −3, −4.8, and −6 dB, respectively. So, we rescale the energy of the intervening speakers in the test set to these SNR levels and evaluate our model performance on different speaker numbers with different SNRs, as shown in Table 6. Increasing the number of intervening speakers without reducing the SNR level will still significantly affect the performance. For example, despite the fact that the training data include a scenario with 3 speakers at −3 dB while the model has not seen any −6 dB mixture at all, it still achieves lower SNR with the 2-speaker test set at −6 dB. In other words, boosting the energy of an intervening speaker by a factor of 4 results in less impact than introducing one extra intervening speaker. The result proves that rescaling noise oversimplifies the complex audio condition in multi-talker scenarios, and illustrates the necessity to study the effect of different speaker numbers.

Table 5 Model performance when a specific expert is activated

Active expert	Performance				
	N_{sp}=1	2	3	4	5
1	**2.05**	4.61	8.90	14.87	19.33
2	2.18	**4.46**	8.53	14.05	18.34
3	2.21	4.48	**8.08**	13.40	17.68
4	2.24	4.59	8.27	**13.30**	**17.25**
All	2.04	4.44	8.17	13.34	17.30

N_{sp} means the speaker number. Best results are in bold

4.3 Scenarios with temporal misalignment

We evaluate the performance of the CMR method under different frame shift numbers, and compare them with several other approaches that handle the misalignment scenario.

Table 6 Model performance under different speaker numbers and different SNR levels

N_{sp}	Performance			
	SNR = 0 dB	−3 dB	−4.8 dB	−6 dB
2	4.44	5.69	6.80	7.67
3	5.81	8.17	10.25	11.35
4	7.38	10.48	13.34	15.66
5	8.10	12.19	15.15	17.30

N_{sp} means the speaker number

1. Unseen: training a model using only perfectly aligned data. This demonstrates the performance degradation in misaligned scenarios when a model has not seen such data during training.

2. Misaligned: directly training with misaligned data as a standard baseline.

3. GT-Aligned: training and evaluating a model with data aligned using the ground-truth frame shift label. This is equivalent to replacing the cross-attention matrix using an oracle matrix with only the respective diagonal set to 1 and all others to 0. In this setting, we assume the actual frame shift number is known during both training and evaluation and serves as an upper bound of methods based on alignment.

4. VoiceFormer: an AVSR conversion of VoiceFormer (Rahimi et al., 2022). This model concatenates audio and video features along the time axis instead of the feature axis.

All models share the same Transformer-based seq2seq structure, and all encoders are initialized with pre-trained parameters.

Table 7 reports the performance of different models under different frame shift numbers. Our method results in low WER under all scenarios and the best overall performance. Although the Unseen model performs the best when the frame shift number is 0, the performance degrades severely even under ±1 frame shift (0.04 s), and becomes unusable as the frame shift grows. This highlights the need to study the task explicitly. VoiceFormer treats audio and video features as separate sequences and only implicitly records their correspondence through positional encoding. Although this prevents the model from being misled by unaligned data, it also makes it more difficult for the model to recognize the alignment information. As a result, it can outperform the Misaligned baseline under large frame shift numbers but falls behind when the frame shift number is low, due to the lack of explicit modeling of cross-modal alignment.

Table 8 WER of models using different auxiliary loss scale

Loss scale	Performance				
	f=−5	−2	0	2	5
0	7.00	5.75	5.39	5.52	5.75
0.01	5.63	4.91	4.76	5.08	5.30
0.1	6.34	5.04	5.13	5.26	5.79

On the contrary, our approach allows the model to retain the input data's alignment information and learn to fix possible misalignment. As a result, it performs significantly better than the Misaligned baseline under almost all scenarios. A highlight of our model is that the model has superior performance when the frame shift number is low, even outperforming the GT-Aligned model in some scenarios. This suggests that the auxiliary task can not only provide the ability to detect misalignment but also allow the model to achieve better feature extraction ability in general. Although the GT-Aligned model benefits from exploiting ground-truth label and performs better with high frame shift numbers, our model can win with stronger generalization when the frame shift number is low and the advantage of using ground-truth becomes less significant.

Table 8 demonstrates the model performance when trained under different auxiliary loss scales. Without the help of auxiliary tasks, the model has difficulty learning the accurate alignment, and the cross attention degrades to an average over the time axis to some extent, which leads to information loss and results in worse performance than the baseline method. This illustrates the necessity of adding the auxiliary task. On the contrary, assigning a large weight to the loss forces the encoder to bias too much toward the auxiliary task, which also slightly reduces performance.

Table 9 shows the effectiveness of masking non-local activation. Fig. 7 visualizes the weights of cross-modal attention of the same test clip at different frame shift values f. Through the auxiliary loss and the attention mask, the model can capture the

Table 7 WER of different methods under different frame shift numbers

Method	Performance											Overall
	f=−5	−4	−3	−2	−1	0	1	2	3	4	5	
Unseen	76.83	69.90	52.47	22.33	6.65	**4.67**	5.89	16.17	46.02	65.86	73.48	40.02
Misaligned	7.01	6.44	5.74	5.17	4.98	5.02	5.04	**4.93**	5.16	5.92	6.16	5.60
GT-Aligned	**5.52**	5.48	5.19	5.15	**4.85**	4.84	5.04	5.02	**5.08**	5.32	**5.24**	5.16
VoiceFormer	6.86	6.27	5.88	5.55	5.18	5.19	5.15	5.50	5.61	5.96	6.08	5.75
CMR	5.63	**5.17**	**5.13**	**4.91**	4.94	4.76	**4.99**	5.08	5.20	**5.20**	5.30	**5.12**

Best results are in bold. The underline means the second-best result

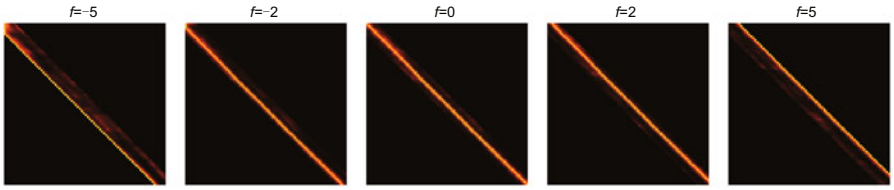

Fig. 7 Visualization of attention weights of the same test clip at different frame shift values f

Table 9 WER of models with or without attention mask

Mode	Performance				
	$f=-5$	-2	0	2	5
With mask	5.98	4.75	4.71	4.97	5.37
Without mask	7.74	5.60	5.19	5.60	6.23

correct alignment and activate the corresponding diagonals, as shown in the figures.

4.4 Diverse multi-talker scenarios

In this subsection, we present overall evaluations of our framework across diverse multi-talker scenarios. Models are trained on data encompassing varying speaker counts and temporal misalignments, and are then evaluated on three specialized test sets. Test set FULL is the full test set containing data mixtures of 1–5 speeches with frame shift numbers ranging between −5 and 5, set SPEAKER is the subset focusing on speaker number variability and only contains data aligned perfectly with different speaker numbers, and set MISALIGNMENT is the subset focusing on misalignment that only contains 2-speaker mixtures with various frame shift numbers.

As shown in Table 10, our approach significantly outperforms the AV-HuBERT backbone (Shi et al., 2022b) across all test conditions. Crucially, although GILA (Hu YC et al., 2023) explicitly models misalignment, its inability to leverage pre-trained representations results in substantially degraded performance. Our integrated solution fully leverages pre-trained model capabilities and achieves further performance gains through explicitly modeling scenario-specific variations.

5 Conclusions

In this work, we address critical challenges in multi-talker AVSR under complex real-world conditions and advance AVSR toward practical deployment in unconstrained settings where speaker numbers are unknown and modality synchronization is imperfect. Our primary contributions form a comprehensive framework for robust AVSR. To explicitly model scenario differences across varying speaker counts, we design the SA-MoE encoder that dynamically assigns processing pathways according to the speaker number. This architecture preserves task-specific feature representations while accommodating characteristic variations across scenarios. For handling temporal misalignment between audio–visual streams, we introduce a novel CMR module that automatically corrects input asynchrony. Integrated seamlessly with pre-trained models, this component employs cross-attention with auxiliary objectives to reconstruct aligned features while enhancing representation learning. Recognizing the inherent difficulty hierarchy across speaker counts, we propose the CBCL training strategy, where the proportion of simpler scenarios decreases exponentially during training. This forces models to progressively focus on challenging multi-speaker cases, optimizing overall performance. All proposed methods are compatible with pre-trained audio–visual encoders and can use knowledge learned from unlabeled data. Experimental validation demonstrates that the SA-MoE effectively captures scenario-specific characteristics across speaker counts, the CMR module significantly outperforms alternatives in asynchronous settings, and CBCL yields consistent gains over standard curriculum approaches.

Despite these advancements, we acknowledge that our work, like most research in audio–visual speech processing, operates within the inherent gap between controlled experimental settings and fully unconstrained real-world conditions. The current study relies on standardized datasets and synthetic misalignment simulations—common practices in the field that nevertheless simplify the complexity of in-the-wild scenarios. Specifically, our framework assumes speaker number estimation can be reasonably obtained, mirrors prior work in using

Table 10 WER of models under different datasets

Method	Performance		
	FULL	SPEAKER	MISALIGNMENT
AV-HuBERT (Shi et al., 2022b)	12.37±0.07	11.51±0.18	6.05±0.09
VoiceFormer (Rahimi et al., 2022)	12.56±0.18	11.70±0.12	5.64±0.32
GILA (Hu YC et al., 2023)	15.15±0.12	13.87±0.07	7.10±0.23
Ours	**11.33±0.12**	**11.12±0.14**	**4.90±0.17**

Best results are in bold

fixed offsets to model audio–visual asynchrony, and evaluates generalization primarily within domain-internal benchmarks. These methodological choices align with contemporary research paradigms but naturally leave open questions about performance under more extreme or dynamic environmental conditions.

These limitations delineate clear and valuable pathways for future work. Moving beyond synthetic simulations to incorporate dynamically shifting misalignments and more robust speaker counting in high-noise environments represents a crucial next step. Furthermore, extending evaluation to diverse, real-world datasets and exploring domain-agnostic adaptation techniques will be essential to bridge the simulation-to-reality gap. We believe our proposed components—designed with compatibility and extensibility in mind—provide a solid foundation for such future explorations toward more robust and deployable audio–visual speech systems.

Contributors

Yuxiao LIN conceived and designed the study, conducted the experiments, and drafted the paper. Tao JIN provided expert guidance throughout the experimental phase and revised the paper. Xize CHENG assisted with data processing and contributed to structuring the paper. Zhou ZHAO and Fei WU supervised the research, provided critical direction, and revised the paper. Yuxiao LIN finalized the paper.

Conflict of interest

Fei WU is an executive associate editor-in-chief of *Frontiers of Information Technology & Electronic Engineering*; he is not involved with the peer review process of this paper. All the authors declare that they have no conflict of interest.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Afouras T, Chung JS, Senior A, et al., 2018a. Deep audio-visual speech recognition. *IEEE Trans Patt Anal Mach Intell*, 44(12):8717-8727. https://doi.org/10.1109/TPAMI.2018.2889052

Afouras T, Chung JS, Zisserman A, 2018b. LRS3-TED: a large-scale dataset for visual speech recognition. https://arxiv.org/abs/1809.00496

Bengio Y, Louradour J, Collobert R, et al., 2009. Curriculum learning. Proc 26th Annual Int Conf on Machine Learning, p.41-48. https://doi.org/10.1145/1553374.1553380

Cheng HY, Liu ZY, Wu W, et al., 2023. Filter-recovery network for multi-speaker audio-visual speech separation. Proc 11th Int Conf on Learning Representations.

Cheng XZ, Jin T, Li LJ, et al., 2023. OpenSR: open-modality speech recognition via maintaining multi-modality alignment. Proc 61st Annual Meeting of the Association for Computational Linguistics, p.6592-6607. https://doi.org/10.18653/v1/2023.acl-long.363

Chung JS, Senior A, Vinyals O, et al., 2017. Lip reading sentences in the wild. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.6447-6456. https://doi.org/10.1109/CVPR.2017.367

Ebbinghaus H, 1885. Über Das Gedächtnis: Untersuchungen Zur Experimentellen Psychologie. Duncker & Humblot, Leipzig, Germany (in German).

Gao RH, Grauman K, 2021. VisualVoice: audio-visual speech separation with cross-modal consistency. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.15490-15500. https://doi.org/10.1109/CVPR46437.2021.01524

Graves A, Bellemare MG, Menick J, et al., 2017. Automated curriculum learning for neural networks. Proc 34th Int Conf on Machine Learning, p.1311-1320.

Gulati A, Qin J, Chiu CC, et al., 2020. Conformer: convolution-augmented Transformer for speech recognition. Proc 21st Annual Conf of the Int Speech Communication Association, p.5036-5040.

Guo PC, Boyer F, Chang XK, et al., 2021. Recent developments on ESPnet toolkit boosted by Conformer. Proc IEEE Int Conf on Acoustics, Speech and Signal Processing, p.5874-5878. https://doi.org/10.1109/ICASSP39728.2021.9414858

Hacohen G, Weinshall D, 2019. On the power of curriculum learning in training deep networks. Proc 36th Int Conf on Machine Learning, p.2535-2544.

Haliassos A, Ma PC, Mira R, et al., 2023. Jointly learning visual and auditory speech representations from raw data. Proc 11th Int Conf on Learning Representations.

Hershey JR, Chen Z, Le Roux J, et al., 2016. Deep clustering: discriminative embeddings for segmentation and separation. Proc IEEE Int Conf on Acoustics, Speech

and Signal Processing, p.31-35.
https://doi.org/10.1109/ICASSP.2016.7471631

Hu EJ, Shen YL, Wallis P, et al., 2022. LoRA: low-rank adaptation of large language models. Proc 10th Int Conf on Learning Representations.

Hu YC, Li RZ, Chen C, et al., 2023. Cross-modal global interaction and local alignment for audio-visual speech recognition. Proc 32nd Int Joint Conf on Artificial Intelligence, p.5076-5084.

Kumar G, Foster G, Cherry C, et al., 2019. Reinforcement learning based curriculum optimization for neural machine translation. Proc Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, p.2054-2061.
https://doi.org/10.18653/v1/N19-1208

Kumar MP, Packer B, Koller D, 2010. Self-paced learning for latent variable models. Proc 24th Int Conf on Neural Information Processing Systems, p.1189-1197.

Kwon Y, Chung SW, 2023. Mole: mixture of language experts for multi-lingual automatic speech recognition. Proc IEEE Int Conf on Acoustics, Speech and Signal Processing, p.1-5.
https://doi.org/10.1109/ICASSP49357.2023.10096227

Li JH, Li CD, Wu YF, et al., 2024. Unified cross-modal attention: robust audio-visual speech recognition and beyond. *IEEE/ACM Trans Audio Speech Lang Process*, 32:1941-1953.
https://doi.org/10.1109/TASLP.2024.3375641

Lin Y, Jin T, Cheng X, et al., 2025. Curriculum learning aided audio-visual speech recognition with arbitrary speaker number. IEEE Int Conf on Acoustics, Speech and Signal Processing, p.1-5.
https://doi.org/10.1109/ICASSP49660.2025.10887654

Ma PC, Petridis S, Pantic M, 2021. End-to-end audio-visual speech recognition with Conformers. Proc IEEE Int Conf on Acoustics, Speech and Signal Processing, p.7613-7617.
https://doi.org/10.1109/ICASSP39728.2021.9414567

Ma PC, Haliassos A, Fernandez-Lopez A, et al., 2023. Auto-AVSR: audio-visual speech recognition with automatic labels. Proc IEEE Int Conf on Acoustics, Speech and Signal Processing, p.1-5.
https://doi.org/10.1109/ICASSP49357.2023.10096889

Makino T, Liao H, Assael Y, et al., 2019. Recurrent neural network transducer for audio-visual speech recognition. Proc IEEE Automatic Speech Recognition and Understanding Workshop, p.905-912.
https://doi.org/10.1109/ASRU46091.2019.9004036

Murre JMJ, Dros J, 2015. Replication and analysis of Ebbinghaus' forgetting curve. *PLoS ONE*, 10(7):e0120644.
https://doi.org/10.1371/journal.pone.0120644

Nagrani A, Chung JS, Zisserman A, 2017. VoxCeleb: a large-scale speaker identification dataset. Proc 18th Annual Conf of the Int Speech Communication Association, p.2616-2620.

Nguyen TS, Stüker S, Waibel A, 2021. Super-human performance in online low-latency recognition of conversational speech. Proc 22nd Annual Conf of the Int Speech Communication Association, p.1762-1766.

Penha G, Hauff C, 2020. Curriculum learning strategies for IR: an empirical study on conversation response ranking. Proc 42nd European Conf on IR Research on Advances in Information Retrieval, p.699-713.
https://doi.org/10.1007/978-3-030-45439-5_46

Petridis S, Stafylakis T, Ma PC, et al., 2018. Audio-visual speech recognition with a hybrid CTC/attention architecture. Proc IEEE Spoken Language Technology Workshop, p.513-520.
https://doi.org/10.1109/SLT.2018.8639643

Platanios EA, Stretcu O, Neubig G, et al., 2019. Competence-based curriculum learning for neural machine translation. Proc Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, p.1162-1172.
https://doi.org/10.18653/v1/N19-1119

Rahimi A, Afouras T, Zisserman A, 2022. Reading to listen at the cocktail party: multi-modal speech separation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.10483-10492.
https://doi.org/10.1109/CVPR52688.2022.01024

Shi BW, Hsu WN, Lakhotia K, et al., 2022a. Learning audio-visual speech representation by masked multimodal cluster prediction. Proc 10th Int Conf on Learning Representations.

Shi BW, Hsu WN, Mohamed A, 2022b. Robust self-supervised audio-visual speech recognition. Proc 23rd Annual Conf of the Int Speech Communication Association, p.2118-2122.

Spitkovsky VI, Alshawi H, Jurafsky D, 2010. From baby steps to leapfrog: how "less is more" in unsupervised dependency parsing. Proc Annual Conf of the North American Chapter of the Association for Computational Linguistics, p.751-759.

Wang WX, Ma GD, Li YK, et al., 2023. Language-routing mixture of experts for multilingual and code-switching speech recognition. Proc 24th Annual Conf of the Int Speech Communication Association, p.1389-1393.

Weinshall D, Cohen G, Amir D, 2018. Curriculum learning by transfer learning: theory and experiments with deep networks. Proc 35th Int Conf on Machine Learning, p.5235-5243.

Xiong W, Droppo J, Huang X, et al., 2016. Achieving human parity in conversational speech recognition.
https://arxiv.org/abs/1610.05256

Yang S, Zhang YH, Feng DL, et al., 2019. LRW-1000: a naturally-distributed large-scale benchmark for lip reading in the wild. Proc 14th IEEE Int Conf on Automatic Face & Gesture Recognition, p.1-8.
https://doi.org/10.1109/FG.2019.8756582

Yang XD, Cheng XZ, Duan JQ, et al., 2024. AudioVSR: enhancing video speech recognition with audio data. Proc Conf on Empirical Methods in Natural Language Processing, p.15352-15361.
https://doi.org/10.18653/v1/2024.emnlp-main.858

Yu D, Kolbæk M, Tan ZH, et al., 2017. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. Proc IEEE Int Conf on Acoustics, Speech and Signal Processing, p.241-245.
https://doi.org/10.1109/ICASSP.2017.7952154

Zhu QS, Zhou L, Zhang ZQ, et al., 2024. VatLM: visual-audio-text pre-training with unified masked prediction for speech representation learning. *IEEE Trans Multimedia*, 26:1055-1064.
https://doi.org/10.1109/TMM.2023.3275873