AMG-AVSR: Adaptive Modality Guidance for Audio-Visual Speech Recognition via Progressive Feature Enhancement

Zhishuo Zhao[*] ZHAOZHISHUO@STU.SCU.EDU.CN
Dongyue Guo DONGYUEGUO@SCU.EDU.CN
Wenjie Ou OUWENJIE@STU.SCU.EDU.CN
Hong Liu[†] LIUHONG@SCU.EDU.CN
Yi Lin[‡] YILIN@SCU.EDU.CN
Sichuan University

Editors: Vu Nguyen and Hsuan-Tien Lin

Abstract

Audio-Visual Speech Recognition (AVSR) is a task that identifies spoken words by analyzing both lip movements and auditory signals. Compared to Automatic Speech Recognition (ASR), AVSR demonstrates greater robustness in noisy environments due to the support of dual modalities. However, the inherent differences between these modalities present a challenge: effectively accounting for their disparities and leveraging their complementary information to extract useful information for AVSR. To address this, we propose the AMG-AVSR model, which utilizes a two-stage curriculum learning strategy and incorporates a feature compression and recovery mechanism. By leveraging the characteristics of different modalities in various scenarios to guide each other, the model extracts refined features from audio-visual data, thereby enhancing recognition performance in both clean and noisy environments. Compared to the baseline model AV-HuBERT, AMG-AVSR demonstrates superior performance on the LRS2 dataset in both noisy and clean environments. AMG-AVSR achieves a word error rate (WER) of 2.9% under clean speech conditions. In various noisy conditions, AMG-AVSR shows a significant reduction in WER compared to previous methods.

Keywords: AV-HuBERT, AVSR, Compression and recovery, Curriculum learning

1. Introduction

With the continuous advancement of neural models (Hinton et al., 2012; Amodei et al., 2016), the performance of automatic speech recognition (ASR) systems has significantly improved, reaching human parity (Xiong et al., 2016) and even surpassing humans in several clean speech benchmarks (Amodei et al., 2016; Tüske et al., 2020). However, ASR systems are highly sensitive to noise, and their performance can degrade drastically when speech recordings are contaminated with noise (Watanabe et al., 2020). Audio-Visual Speech Recognition (AVSR) methods combine audio and video modalities, leveraging noise-invariant lip movement information to make AI systems closer to human speech perception (McGurk and MacDonald, 1976). Recently, new model architectures (Afouras et al., 2018a; Ren et al.,

[*] First author.
[†] Corresponding author 1.
[‡] Corresponding author 2.

2021) and large-scale data collection (Afouras et al., 2018c; Makino et al., 2019) have made significant progress in AVSR tasks.

However, there are inherent differences between audio and visual modalities. For instance, the audio modality typically contains more speech information but is susceptible to noise interference (Watanabe et al., 2020), while the video modality provides visual cues such as lip movements (Petridis et al., 2018). In clean environments, ASR accuracy is generally higher than Visual Speech Recognition (VSR), whereas VSR exhibits better noise resistance in noisy conditions (Kinoshita et al., 2021), which aligns with the biological observation that humans rely more on auditory cues in clean environments and visual cues in noisy environments (Li et al., 2023). Based on these differences, can we use the more reliable modality under different conditions to guide the less reliable modality in learning relevant knowledge, thereby obtaining enhanced information from both the audio and visual modalities to achieve robust AVSR?

Based on this concept, we propose AMG-AVSR (Adaptive Modality Guidance for Audio-Visual Speech Recognition), a model that incorporates a Multi-Scale Compression and Recovery (CAR) module, along with two types of effective fine-tuning curriculum learning strategies. AMG-AVSR is based on the pre-trained audio-visual aligned encoder AV-HuBERT (Shi et al., 2022a), co-trained on a large number of multi-modality high-resource domain utterances, to align different modalities in the same phoneme space using the same encoder, making cross-modal guidance possible under different conditions.

We integrated multi-scale CAR modules into the pre-trained Transformer encoder, allowing AMG-AVSR to adaptively learn refined information from the redundancy in the audio-visual feature space, thus enhancing feature refinement. Furthermore, the two curriculum learning strategies sequentially guide the model through different modalities. Curriculum learning for Modal (C-Modal) leverages the rich text-related knowledge from the audio modality to guide the visual modality for joint text information mapping, enabling AMG-AVSR to learn from unimodal tasks to multimodal recognition, thereby enhancing training effectiveness. Considering the superior noise resistance of visual information in noisy environments (Kinoshita et al., 2021), the C-Noise strategy transitions from clean AVSR to noisy AVSR, using the visual modality to guide the audio modality under noise, gradually adapting AMG-AVSR to noise and enhancing its robustness.

The main contributions are as follows:

- To the best of our knowledge, AMG-AVSR is the first model to achieve mutual guidance learning between audio and video modalities under different conditions, fully considering the different recognition characteristics of each modality.

- The Multi-Scale CAR modules in AMG-AVSR utilize compression and recovery of features within the feature space, enhancing the features and effectively extracting refined text-related information, which is key to enhancing information from different modalities.

- We investigate the performance and robustness of the proposed approach on the LRS2 dataset. AMG-AVSR achieves a word error rate (WER) of 2.9% in clean speech conditions and significantly reduces WER in noisy conditions compared to previous methods.

2. Related Work

2.1. Audio-Visual Speech Recognition

With the advent of deep learning technologies and multimodal data processing frameworks, AVSR systems have continuously evolved, demonstrating improved performance. The key research direction remained how to better filter noise and extract critical speech information in noisy environments. For example, Huang et al. (2023) addressed the challenges of speech recognition in noisy environments by employing multi-layer cross-attention mechanisms. Ma et al. (2023) significantly improved recognition performance in noisy environments through self-supervised learning and automatic label generation. GILA (Hu et al., 2023) captured deeper associations between audio and visual features by introducing global interaction and local alignment, thereby enhancing AVSR accuracy and robustness.

Unsupervised learning also made significant strides in the AVSR field. AV-HuBERT(Shi et al., 2022b) , through pre-training on large amounts of unlabeled audio-visual datasets, generated fine-grained feature representations, which could be fine-tuned for various downstream tasks such as VSR and AVSR. AV2vec(Zhang et al., 2023) further enhanced multimodal data alignment using multi-layer convolutional networks and Transformer architectures. AV-wav2vec (Zhu et al., 2023) utilized multi-channel audio data for self-supervised training, further improving the robustness and accuracy of AVSR.

For fine-tuning on downstream tasks such as AVSR, Shi et al. (2022c) removed the classification linear layer and added a decoder for sequence-to-sequence (Seq2Seq) fine-tuning. Ren et al. (2023) replaced the Transformer with a Conformer and modified the audio-visual front-end for fine-tuning. Most existing studies fine-tune the output of pre-trained models but do not address the handling of redundant information in feature dimensions.

To effectively fine-tune pre-trained models for robust AVSR tasks, we propose a Multi-Scale CAR framework based on audio-visual pretraining methods. This framework focuses on extracting refined text-related information and enhancing the original features during the model fine-tuning stage.

2.2. Curriculum Learning

Curriculum learning, proposed by Bengio et al. (2009), involves progressively increasing task difficulty to enhance a model's learning efficiency and generalization. This approach starts with simple tasks and gradually introduces more complex ones. In recognition tasks, curriculum learning is crucial. For instance, Wang et al. (2017) employed curriculum learning to train an end-to-end speech recognition model by starting with short sentences and increasing their length. Kim et al. (2024) used audio-visual speech units and progressively reduced reliance on audio, eventually training solely with visual speech units, enhancing visual speech recognition capabilities.

Inspired by these methods, we propose two curriculum learning strategies: C-Modal and C-Noise. C-Modal transitions from audio recognition to audio-visual recognition, using the information-rich audio modality to assist the visual modality mapping, enabling the model to learn from simple unimodal tasks to complex multimodal tasks. C-Noise transitions from clean AVSR to noisy AVSR, using the noise-resistant visual modality to guide the audio

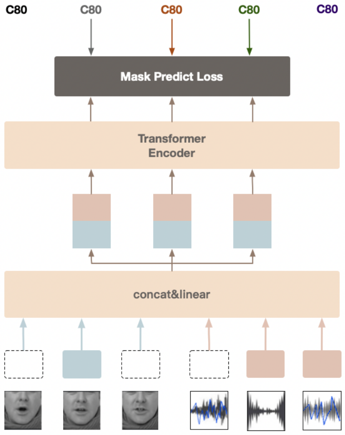


Figure 1: AV-HuBERT pretrain for Audio-Visual Speech Recognition. Black waveform: original audio; Blue waveform: noise; Cn: audio-visual clusters; Dashed box: Masked information.

modality under noisy conditions, gradually adapting the model to noise and enhancing its robustness.

3. METHOD

3.1. AV-HuBERT Pretrain

AV-HuBERT (Shi et al., 2022b) is a self-supervised representation learning method for audio-visual speech, as shown in Figure 1. The model operates in two primary phases: feature clustering and mask prediction. During the feature clustering phase, a discrete latent variable model (e.g., k-means) assigns frame-level labels to the audio-visual speech data. This process involves extracting image sequences $V = \{V_t\}_{t=1}^{T}$ and audio acoustic frames $A = \{A_t\}_{t=1}^{T}$, which can be either Mel-frequency cepstral coefficients (MFCC) or audio-visual features from a previous encoder. These features are then used to generate sequences $z = \{z_{ta}\}_{t=1}^{T}$, representing the clustered assignments.

The resulting paired data (A, V, z) is subsequently used to train the model in the mask prediction phase. In the mask prediction phase, a mask is applied to portions of the input sequences, and the model is trained to predict the masked parts. This process is analogous to BERT(Kenton and Toutanova, 2019)'s masked language modeling, where the goal is to learn improved audio-visual representations in the speech space $f_p = \{f_{tp}\}_{t=1}^{T} \in \mathbb{R}^{T \times D}$, where T is the sequence length and D is the embedding dimension. The model alternates between these two phases, and with each iteration, it improves the quality of the audio-visual speech clustering and representations. Through this iterative approach, AV-HuBERT not only enhances its ability to capture nuanced speech features but also allows different modalities to be projected into the same phoneme space, achieving cross-modal alignment.AMG-AVSR utilizes this joint mapping mechanism to enable mutual guidance between the two modalities, refining and enhancing the feature information.

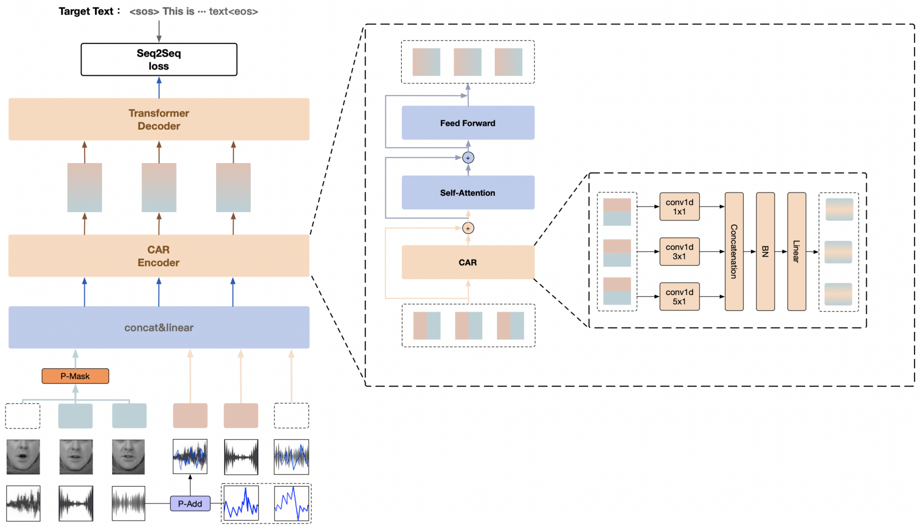


Figure 2: AMG-AVSR finetune for Audio-Visual Speech Recognition. Black waveform: original audio; Blue waveform: noise; Dashed box: CAR Encoder part; P-Add: Progressive noise addition module; P-Mask: Progressive unmasking video information module.

3.2. Finetuning For AVSR

3.2.1. Overall Architecture

Figure 2 illustrates AMG-AVSR, which builds upon the pre-trained AV-HuBERT model (Shi et al., 2022b). To further refine the features extracted by the encoder during the fine-tuning phase, we retain the pre-trained AV-Fusion module and encoder, while integrating multi-scale CAR modules into each layer, forming the CAR Encoder. We remove the cluster prediction head of the pre-trained model and add randomly initialized Transformer decoder blocks after the pre-trained encoder. Additionally, we incorporate a Progressive unmasking video information module (P-Mask) and a progressive noise addition module (P-Add) into the audio extraction module to achieve our two curriculum learning objectives: a modality-specific curriculum learning strategy (C-Modal) and a noise-specific curriculum learning strategy (C-Noise). Together with the multi-scale CAR modules, these enhancements use the more reliable modality in different scenarios to guide the less reliable modality for mapping learning, and perform Seq2seq loss in the combined text space through the decoder. The overall model architecture is shown in Figure 2.

3.2.2. CAR Module

Existing AVSR methods face significant challenges in handling multimodal features, particularly in effectively capturing and processing critical features within the combined audio-visual information (Ivanko et al., 2023). To address these limitations, we propose the Multi-Scale CAR module. Unlike traditional feature extraction methods, such as the Squeeze-and-Excitation (SE) block (Hu et al., 2018), which uses global average pooling, the Multi-Scale CAR module focuses on the relationships between different feature granularities across various channels to extract more refined information representations.

As shown in Figure 2, we add the CAR_j module to each Layer_j of the Transformer encoder. Utilizing this module, the information passed through the encoder undergoes compression and recovery at different feature granularities, capturing feature relationships at various scales. First, the input $x_j = \{x_t^j\}_{t=1}^{T} \in \mathbb{R}^{T \times D}$ of each layer of the encoder is processed by the CAR_j module, which includes three 1D convolutional layers with kernel sizes $i = 3, 5,$ and 7. These convolutional layers reduce the original feature dimensions of the encoder's input dimension e to a lower feature dimension c at different granularities.

$$F_i = \text{Conv1D}_i(X_j) \quad (\text{kernel\_size} = (3, 5, 7)\ \text{output\_dimension} = c) \tag{1}$$

The outputs of the three convolutional layers are concatenated along the channel dimension, forming a feature matrix that contains multi-scale information, and passed through a Batch Normalization (BN) layer to normalize the feature distribution:

$$F_{\text{BN}} = \text{BatchNorm}(F_3 \oplus F_5 \oplus F_7) \tag{2}$$

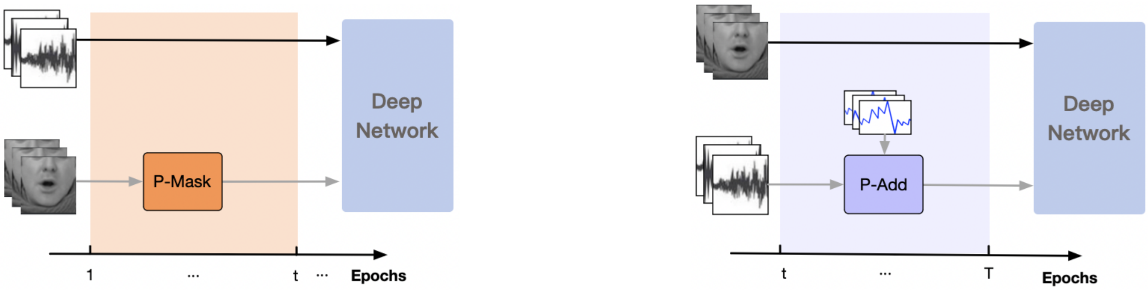
Next, a linear layer is used to project the compressed information back to the original dimension and update X_j:

$$X_j' = \text{Linear}(F_{\text{BN}}) + X_j \tag{3}$$

Subsequently, the updated representation X_j' is passed into the remaining modules of the encoder to maintain the learned syntactic knowledge:

$$X_{j+1} = \text{Layer}_j(X_j') \tag{4}$$

3.3. Curriculum Learning



(a) C-Modal.

(b) C-Noise.

Figure 3: The curriculum learning strategies corresponding to the training phases of AMG-AVSR. (a) C-Modal curriculum learning strategy (b) C-Noise curriculum learning strategy; DeepNetwork: the backend encoder and decoder structure of AMG-AVSR.

The audio modality contains more speech information but is susceptible to noise, while the video modality provides visual cues like lip movements. To leverage these differences, we propose C-Modal and C-Noise curriculum learning strategies. As shown in the figure 3, these strategies run sequentially. C-Modal transitions from unimodal to multimodal training, using clean audio to guide visual information, improving AVSR in clean conditions. C-Noise starts with clean AVSR tasks and progressively introduces noise, using visual information to help the model exclude noise and enhance robustness.

3.3.1. Curriculum Learning for Modal (C-Modal)

To leverage audio information to guide visual representations and enable the model to better learn modality representations and knowledge from simple to complex, we propose the C-Modal method. As shown in the C-Modal phase in Figure 3, we incorporate a P-Mask architecture in the video front-end to gradually control the probability p_{modal} of masking the video information, transitioning the training from audio input to audio-visual input, and ultimately using complete audio-visual information.

The embeddings of the visual features $f_v \in \mathbb{R}^{T\times D}$ and audio features $f_a \in \mathbb{R}^{T\times D}$ are concatenated as $g = f_a \oplus M(f_v)$, where T is the sequence length, D is the dimension of embedding, and $\oplus$ represents the concatenation operation in the embedding dimension. Therefore, $g \in \mathbb{R}^{T\times 2D}$ is the concatenated feature. $M(\cdot)$ is a masking function that randomly masks out p_{model} of frames from the input sequence.

Initially, we train the decoder with only audio input, which is more efficient due to the smaller data volume compared to simultaneous audio-visual input. As training progresses, we unfreeze the encoder and gradually reduce p_{model} from 1 to 0. This method allows the audio stream, which contains more detailed speech information, to guide the visual modality through the CAR encoder for joint mapping. This process helps the visual modality obtain more refined representations, thereby enhancing AVSR performance.

3.3.2. Curriculum Learning for Noise (C-Noise)

During the pre-training phase of AV-HuBERT, noise augmentation is achieved by randomly adding different types of noise to the audio input (Shi et al., 2022c). To further improve the model's noise robustness, we adopt the C-Noise strategy following the C-Modal strategy during the fine-tuning phase. In the C-Modal phase, audio features have already guided the visual features to obtain refined representations. Therefore, in the subsequent C-Noise phase, we can use visual features to guide the noisy audio features to extract speech-related representations, thereby enhancing the model's robustness in noisy environments.

As shown in the C-Noise phase in Figure 3. We use the P-Add module to progressively add noise to obtain noisy audio features $f_t^{a_{\text{noise}}}$. Specifically, the noise-augmented audio features are represented as:

$$f_{\text{a}}^{\text{noise}} = \begin{cases} f_a^{\text{clean}}(t) & \text{with probability } 1 - p_{\text{noise}}(t) \\ f_a^{\text{clean}}(t) + \text{noise}(t) & \text{with probability } p_{\text{noise}}(t) \end{cases} \quad \forall 0 \le t \le T \tag{5}$$

where noise(t) is the noise extracted from the MUSAN dataset, including "natural", "music", "babble", and "speech". The probability $p_{\text{noise}}(t)$ represents the probability of adding noise, and during training, $p_{\text{noise}}(t)$ gradually increases from 0 to 1.

Next, the adjusted audio features is concatenated with the original video features:

$$g_{\text{noise}} = f_a^{\text{noise}} \oplus f_v^{\text{clean}} \tag{6}$$

4. EXPERIMENT

4.1. Data and Experimental Setup

Our experiments utilize the LRS2 dataset (Afouras et al., 2018a), which includes around 224 hours of audio-visual speech from more than 1,000 speakers, making it one of the most extensive publicly available labeled datasets for Audio-Visual Speech Recognition. This dataset features a wide variety of utterances from British English television broadcasts. We follow the pre-processing steps described in (Shi et al., 2022a) to segment and align the audio-visual data, ensuring high-quality synchronization between the audio and video streams. The original dataset divides the training data into two parts: Pretrain (195 hours) and Train (29 hours), both transcribed from videos to text at the sentence level. The main difference is that the video clips in the Pretrain partition are not strictly trimmed and are sometimes longer than the corresponding text. We conduct experiments on LRS2 using different amounts of training data (Pretrain+Train (224 hours) and Train (29 hours)).

Additionally, we enhance input samples using various noise categories. The noise audio clips for the "natural," "music," and "babble" categories are sourced from the MUSAN dataset (Snyder et al., 2015), while the overlapping "speech" noise samples come from the LRS2 dataset itself. When creating the "speech" and "babble" noise sets, we ensure there are no speaker overlaps between different partitions.

For all our experiments, we use the AV-HuBERT LARGE architecture as the default model. This model consists of 24 transformer blocks, each with 16 attention heads and 1024/4096 embedding/feedforward dimensions. During finetuning, we add a 9-layer transformer decoder with similar embedding/feedforward dimensions, initialized randomly.Further experimental details can be found in the appendix.

4.2. Evaluation and Implementation Details

In all our experiments on the LRS2 dataset, we use the word error rate (WER) as the evaluation metric for AVSR. The WER is calculated using the formula:

$$\text{WER} = \frac{S + D + I}{M} \times 100\% \tag{7}$$

where S represents the number of substitutions, D represents the number of deletions, I represents the number of insertions, and M is the total number of words in the reference.

To ensure robustness and reliability, we perform multiple runs for each experiment and report the average WER. Additionally, we adhere to the standard data preprocessing and augmentation techniques described in Shi et al. (2022a), including video frame sampling, audio normalization, and the addition of various noise types during training.

During training, we add different types of noise (natural, music, babble, and speech) to the audio-visual samples and gradually increase the probability and intensity of the noise. For evaluation, we test the model under various noise conditions, including clean audio and noise added at SNR levels of -10, -5, 0, 5, 10dB.

4.3. Main Result

Table 1 shows the amount of labeled audio-visual speech data, dataset details, and corresponding Word Error Rates (WER) for models using the LRS2 dataset. The baseline model

Table 1: WER (%) of AMG-AVSR and previous works on the LRS2 dataset. ”Labeled Utt(hrs)” denotes the amount of labeled audio-visual speech data used in each system (in hours).

Labeled Utt(hrs)	Dataset	Method	WER(%)↓
1391	MV-LRS, LRS2 and LRS3	TM-seq2seq(Afouras et al., 2018a)	8.3
381	LRS2 and LRW	CTC/Attention(Petridis et al., 2018)	7.0
3448	LRW, LRS2, LRS3, VoxCeleb2 and AVSpeech	AUTO-AVSR (Ma et al., 2023)	**1.5**
224	LRS2	DCM(Lee et al., 2020)	8.6
	LRS2	TDNN(Yu et al., 2020)	5.9
	LRS2	Hyb-Conformer(Ma et al., 2021b)	3.7
	LRS2	AV-HuBERT(Shi et al., 2022c)	3.1
	LRS2	GILA (Hu et al., 2023)	3.1
	LRS2	**AMG-AVSR(ours)**	<u>2.9</u>
29	LRS2 (Train)	AV-HuBERT(Shi et al., 2022c)	5.1
	LRS2 (Train)	**AMG-AVSR(ours)**	4.2

Table 2: WER (%) of AMG-AVSR and AV-HuBERT on the LRS2 dataset. “Mode” denotes whether a model uses audio-visual input (AV) or only audio as input (A). “Hr” denotes the amount of labeled audio-visual speech data used in each system.

Model	Mode	Hr	Babble, SNR=					Speech, SNR=					Music+Natural, SNR=					Clean
			-10	-5	0	5	10	-10	-5	0	5	10	-10	-5	0	5	10	
AV-HuBERT	AV	29	35.8	17.5	10.8	8.9	5.9	11.9	7.8	6.8	6.1	5.4	15.6	9.5	8.0	6.2	5.6	<u>5.2</u>
AV-HuBERT	AV	224	32.6	15.2	9.2	6.4	3.9	9.1	6.1	4.2	3.6	3.4	13.1	6.9	5.5	3.7	3.5	<u>3.1</u>
AMG-AVSR(ours)	A	29	98.1	76.2	29.6	15.9	9.4	81.1	56.4	31.8	19.5	10.1	48.3	29.3	12.9	9.6	7.3	6.5
		224	99.6	69.2	20.6	9.94	5.2	74.2	51.2	27.4	16.7	8.5	43.7	25.1	9.9	6.7	4.2	3.2
AMG-AVSR(ours)	AV	29	31.2	14.5	8.9	6.5	5.0	10.6	7.1	6.2	5.1	4.7	14.4	9.2	7.2	5.3	4.6	**4.2**
		224	30.6	13.4	7.7	5.2	3.8	8.7	5.6	3.9	3.5	3.4	12.5	6.8	5.5	3.6	3.3	**2.9**

(Shi et al., 2022c) achieved a WER of 5.1% with 29 hours of labeled data and 3.1% with 224 hours. In contrast, our proposed model achieved a WER of 4.2% with 29 hours of labeled data, significantly outperforming the baseline model and many other models trained with more data. For instance, Afouras et al. (2018a) used 1391 hours of data to achieve an 8.3% WER, Petridis et al. (2018) used 381 hours to achieve a 7.0% WER, and Yu et al. (2020) used 224 hours to achieve a 5.9% WER.

Furthermore, with 224 hours of labeled data, our model's WER reduces to 2.9%, only higher than AUTO-AVSR's 1.5% WER (Ma et al., 2023). It is important to note that AUTO-AVSR used 3448 hours of labeled data for training, while AMG-AVSR achieved remarkable data efficiency with only 224 hours of training data.

4.4. Analysis

4.4.1. Advantage of Audio-Visual Modalities Compared to Audio-Only Modalities

From the data in the Table 2, the audio-visual (AV) modality significantly outperforms the single audio (A) modality under both clean and noisy conditions. In clean conditions, the

Table 3: Effect of CAR and Curriculum Learning. Different approaches for CAR and Curriculum Learning methods are selected, including modality-specific and noise-specific curriculum learning strategies. ”Clean” represents the clean condition, and ”Babble” represents the condition with babble noise added at SNR=0 at each time step.

Methods		224h WER (%)		29h WER (%)	
		Clean	Babble	Clean	Babble
Baseline (Shi et al., 2022c)		3.143	9.224	5.226	10.848
CAR	+conv3	3.091	8.201	4.474	9.454
	+conv3 +conv5	3.072	8.094	4.369	9.281
	+conv3 +conv5 +conv7	3.072	7.951	4.343	9.106
Curriculum Learning	C-Modal	3.043	9.236	4.452	10.837
	C-Noise	3.127	7.975	5.110	8.982
	C-Modal + C-Noise	3.076	7.887	4.631	8.982
CAR + Curriculum Learning		**2.922**	**7.692**	**4.214**	**8.884**

WER for the audio modality is 3.2% with 224 hours of training and 6.5% with 29 hours, compared to 2.9% and 4.2% for the audio-visual modality, respectively. In noisy conditions, such as Babble noise at SNR=0, the WER for the audio modality is 29.6%, while it is only 8.9% for the audio-visual modality, highlighting its superior robustness to noise. These results show that the audio-visual modality significantly enhances both recognition accuracy in clean environments and robustness in noisy conditions.

4.4.2. **Effective Improvement Compared to the Original AV-HuBERT**

As shown in Table 2 ,AMG-AVSR shows significant performance improvements over AV-HuBERT in both clean and noisy conditions. In clean conditions, with 29 hours of training data, AMG-AVSR achieves a WER of 4.2% compared to AV-HuBERT’s 5.2%. With 224 hours of data, the WER drops to 2.9%, while AV-HuBERT’s is 3.1%. These results demonstrate that AMG-AVSR has better data utilization and higher accuracy in clean environments.

AMG-AVSR demonstrates exceptional performance in noisy conditions, particularly under Babble noise. At SNR = -10 dB, AMG-AVSR achieves a WER of 30.6 %, compared to AV-HuBERT’s 32.6%. At SNR = -5 dB, AMG-AVSR achieves a WER of 13.4%, whereas AV-HuBERT’s WER is 15.2%. These results highlight our model’s superior robustness to noise.

4.5. Ablation Studies

In this section, we investigate the impact of each individual building block by testing them on the LRS2 dataset. Additionally, to more intuitively observe the improvements of AMG-AVSR in noisy conditions, we actively added Babble noise with an SNR of 0 to reflect our model’s performance in noisy environments.

4.5.1. CAR model Contribution in Audio-Visual Speech Recognition

We analyzed the impact of the CAR model on AVSR performance, as shown in Table 3. The table compares the WER (word error rate) of the baseline model (AV-HuBERT) and our proposed model using different convolution scales (conv3, conv5, conv7).

By adding compression with a convolution scale of 3, and using its dimension reduction and recovery operations, the WER for the model trained for 224 hours decreases to 3.091% in clean conditions and 8.201% in noisy conditions. The model trained for 29 hours shows similar improvements.

Further applying parallel compression with convolution scales of 5 and 7, then concatenating and recovering the compressed results, allows the model to learn richer information. The WER for the model trained for 224 hours decreases to 3.072% in clean conditions and 7.951% in noisy conditions. Similarly, the model trained for 29 hours shows a reduction in WER to 9.106% in noisy conditions.

4.5.2. Curriculum Learning Strategy For Clean And Noisy Inputs

We also explored the impact of two curriculum learning strategies on AVSR performance in clean and noisy conditions, as detailed in Table 3. Introducing the Curriculum Learning strategy for Modal (C-Modal) reduced the WER to 3.043% in clean conditions. Similarly, for the model trained for 29 hours, the WER in clean conditions also decreased to 4.452%, but there was no significant decrease in noisy conditions.

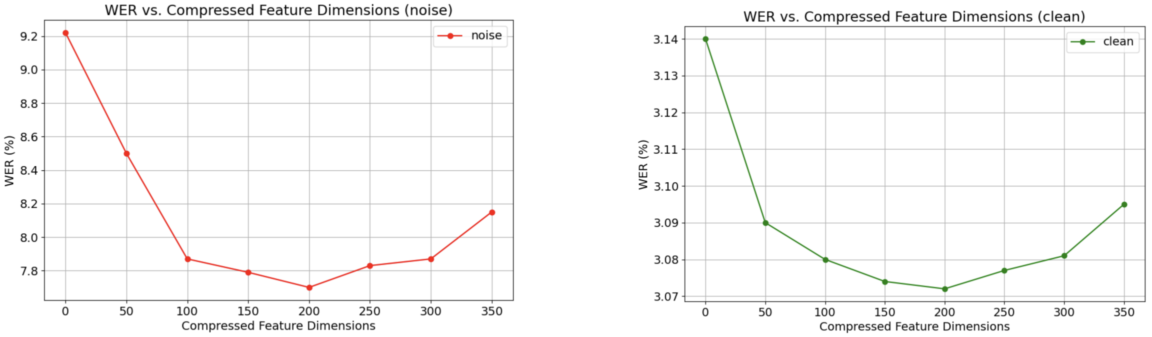
For the curriculum learning strategy that introduces noise (C-Noise), it effectively improved robustness in noisy conditions, reducing the WER to 7.975%. The model trained with the 29-hour dataset also reduced its WER to 8.892%.

Combining these two strategies effectively reduced WER in both clean and noisy conditions. For the model trained with 224-hour data, the WER decreased to 3.076% in clean conditions and to 7.887% in Babble noise conditions. For the model trained with 29-hour data, the WER decreased to 4.631% in clean conditions and to 8.982% in Babble noise conditions.

When combining feature compression and curriculum learning methods, the model trained with 224-hour data shows a further reduction in WER to 2.922% in clean conditions and to 7.692% in noisy conditions. Similarly, the model trained with 29-hour data shows a decrease in WER to 4.214% in clean conditions and to 8.884% in noisy conditions. This indicates that combining feature compression with curriculum learning leverages the advantages of both methods, significantly improving recognition accuracy in various environments.

4.5.3. The Effect of Compressed Feature Dimensions on Recognition Accuracy

As shown in Figure 4, we present the impact of different compression dimensions c on accuracy. It can be seen that as the compression dimension c increases, the model's WER (Word Error Rate) shows significant changes in both clean and noisy speech scenarios. As illustrated in the "Results of babble noise audio" and "Results of clean audio" sections of Figure 4, at a compression dimension of 0, where no compression is applied (our baseline), the WER reaches its highest levels, at 9.2% and 3.14% in clean and noisy speech, respectively. This is because the feature dimensions are not compressed and still contain a lot of redundant and noisy information.



(a) Results of babble noise audio.

(b) Results of clean audio.

Figure 4: Overall results of different compression dimensions on 224 dataset. (a) Results of babble noise audio. (b) Results of clean audio.

As the compression scale increases, the WER continues to decrease for both clean and noisy speech, indicating that within this range, reducing the compression dimension helps the model better capture and represent key information. However, when the compression dimension continues to increase to 350, the WER rises to 8.15% and 3.095%, respectively. This indicates that a smaller compression rate (larger compression dimension) leads to the model's inability to effectively capture key features during learning, thus affecting its ability to extract critical information.

4.5.4. The impact of parameters and training steps on experimental results

To compare the impact of parameters and training steps on experimental results, we replicated AV-HuBERT under the same dataset and initialization conditions. Both AV-HuBERT and AMG-AVSR were trained on the 224-hour and 29-hour datasets for 120k and 60k steps, respectively, and both converged. We evaluated their parameters and WER under clean and noisy (SNR=0, babble noise) conditions. The results show that adding the CAR module slightly increased the model size by 2.07% (from 477.3M to 487.2M) but led to significant WER reductions, nearly 20% under noisy conditions, in both dataset scenarios, demonstrating the effectiveness of the CAR module. Detailed results are shown in Table 4 and Table 5.

Table 4: Model Parameter Comparison Based on the 224-Hour Dataset.

Methods	**Steps**	**Param**	**WER (Clean)**	**WER (Babble)**
		(MB)	**224h**	**224h**
AV-HuBERT	120k	477.3 (×1.00)	3.14	9.22
AMG-AVSR	120k	**487.2** (×1.02)	**2.92**	**7.69**

5. Conclusion

In this study, we proposed the AMG-AVSR model, which integrates Compression and Recovery (CAR) structures with curriculum learning strategies, and we explored its effects

Table 5: Model Parameter Comparison Based on the 29-Hour Dataset.

Methods	**Steps**	**Param** **(MB)**	**WER (Clean)** **29h**	**WER (Babble)** **29h**
AV-HuBERT	60k	477.3 (×1.00)	5.22	10.85
AMG-AVSR	60k	**487.2** (×1.02)	**4.21**	**8.88**

on Audio-Visual Speech Recognition (AVSR) tasks. Our experiments on the LRS2 dataset demonstrated that these methods significantly improved the model's performance in both clean and noisy environments. Leveraging guidance from different modalities and compressing and recovering feature dimensions not only enhanced data utilization efficiency but also increased the model's robustness under various noise conditions. AMG-AVSR exhibited excellent performance with a significant reduction in word error rate (WER), highlighting the importance of these techniques in advancing AVSR technology. These findings emphasize the potential of utilizing differences between modalities to improve the accuracy and reliability of AVSR systems, making them more effective in diverse environments.

Acknowledgments

This work was supported National Natural Science Foundation of China under Grant 62371323 and U2333209.

References

T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. Deep audio-visual speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018a.

T. Afouras, J. S. Chung, and A. Zisserman. The conversation: Deep audio-visual speech enhancement. In *Proceedings of the Annual Conference of the INTERSPEECH*, 2018b.

T. Afouras, J. S. Chung, and A. Zisserman. Lrs3-ted: A large-scale dataset for visual speech recognition. *ArXiv preprint arXiv:1809.00496*, 2018c.

D. Amodei et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48. ACM, 2009.

J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. *ArXiv preprint arXiv:1806.05622*, 2018.

G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018.

Y. Hu, R. Li, C. Chen, et al. Cross-modal global interaction and local alignment for audio-visual speech recognition. *ArXiv preprint arXiv:2305.09212*, 2023.

C. Huang, J. Chen, and Z. Wu. Cross-modal speech recognition using multi-layer attention. *Computer Speech & Language*, 2023.

Denis Ivanko, Dmitry Ryumin, and Alexey Karpov. A review of recent advances on deep learning methods for audio-visual speech recognition. *Mathematics*, 11(12):2665, 2023.

J. D. M. W. C. Kenton and L. K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, volume 1, page 2, 2019.

J. Kim et al. Pretraining with audio-visual speech units. *ArXiv preprint arXiv:2401.12345*, 2024.

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

K. Kinoshita, T. Ochiai, M. Delcroix, and T. Nakatani. Improving noise robust automatic speech recognition with single-channel time-domain enhancement network. In *Proceedings of the IEEE ICASSP*, 2021.

Yong-Hyeok Lee, Dong-Won Jang, Jae-Bin Kim, Rae-Hong Park, and Hyung-Min Park. Audio–visual speech recognition based on dual cross-modality attentions with the transformer model. *Applied Sciences*, 10(20):7263, 2020.

Dengshi Li, Yu Gao, Chenyi Zhu, Qianrui Wang, and Ruoxi Wang. Improving speech recognition performance in noisy environments by enhancing lip reading accuracy. *Sensors*, 23 (4):2053, 2023. URL https://mdpi-res.com/d_attachment/sensors/sensors-23-02053/article_deploy/sensors-23-02053.pdf.

P. Ma, Z. Yan, and L. Xie. Lip by lip: Unsupervised learning for lip reading. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13078–13086, 2021a.

Pingchuan Ma, Stavros Petridis, and Maja Pantic. End-to-end audio-visual speech recognition with conformers. In *ICASSP 2021-2021 IEEE*, pages 7613–7617. IEEE, 2021b.

Pingchuan Ma, Alexandros Haliassos, Adriana Fernandez-Lopez, Honglie Chen, Stavros Petridis, and Maja Pantic. Auto-avsr: Audio-visual speech recognition with automatic labels. *ArXiv preprint arXiv:2304.01234*, 2023.

T. Makino, B. Lee, and S. Watanabe. Recurrent neural network transducer for audio-visual speech recognition. *ArXiv preprint arXiv:1910.04984*, 2019.

J. Martinez. Lipreading with densenet and resnet. *ArXiv preprint arXiv:2110.00054*, 2021.

H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, 1976.

Wenjie Ou, Zhishuo Zhao, Dongyue Guo, Zheng Zhang, and Yi Lin. Winnet: Make only one convolutional layer effective for time series forecasting. In *Advanced Intelligent Computing Technology and Applications (ICIC)*, 2024. URL https://doi.org/10.1007/978-981-97-5678-0_30.

Stavros Petridis, Themos Stafylakis, Pei Ma, Peng Cai, and Maja Pantic. Audio-visual speech recognition with a hybrid ctc/attention architecture. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 513–520. IEEE, 2018.

S. Ren, Q. Hu, X. Liu, Y. Cheng, and Z. Ren. Audiovisual speech recognition with multimodal recurrent neural networks. *Neural Networks*, 136:59–69, 2021.

X. Ren, C. Li, S. Wang, et al. Practice of the conformer enhanced audio-visual hubert on mandarin and english. In *ICASSP 2023-2023*, pages 1–5. IEEE, 2023.

B. Shi, W. N. Hsu, K. Lakhotia, et al. Learning audio-visual speech representation by masked multimodal cluster prediction. *ArXiv preprint arXiv:2201.02184*, 2022a.

B. Shi, A. Mohamed, and W. N. Hsu. Learning lip-based audio-visual speaker embeddings with av-hubert. *ArXiv preprint arXiv:2205.07180*, 2022b.

Bowen Shi, Wei-Ning Hsu, and Abdelrahman Mohamed. Robust self-supervised audio-visual speech recognition. *arXiv preprint arXiv:2201.01763*, 2022c.

David Snyder, Guoguo Chen, and Daniel Povey. Musan: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*, 2015. URL https://arxiv.org/abs/1510.08484.

Z. Tüske, G. Saon, K. Audhkhasi, and B. Kingsbury. Single headed attention based sequence-to-sequence model for state-of-the-art results on switchboard-300. In *Proceedings of the Annual Conference of the INTERSPEECH*, 2020.

X. Wang et al. End-to-end speech recognition with curriculum learning. In *Proceedings of the Annual Conference of the INTERSPEECH*, 2017.

S. Watanabe, M. Mandel, J. Barker, and E. Vincent. Chime-6 challenge: Tackling multi-speaker speech recognition for unsegmented recordings. *ArXiv preprint arXiv:2004.09249*, 2020.

W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig. Achieving human parity in conversational speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(4):644–656, 2016.

Joon Son Yu, Themos Stafylakis, Pei Ma, Stavros Petridis, and Maja Pantic. Audio-visual recognition of overlapped speech for the lrs2 dataset. In *ICASSP 2020-2020 IEEE*, pages 6984–6988. IEEE, 2020.

J.X. Zhang, G. Wan, Z.H. Ling, et al. Self-supervised audio-visual speech representations learning by multimodal self-distillation. In *ICASSP 2023-2023*, pages 1–5. IEEE, 2023.

Qiushi Zhu, Jie Zhang, Yu Gu, Yuchen Hu, and Lirong Dai. Multichannel av-wav2vec2: A framework for learning multichannel multi-modal speech representation. *arXiv preprint arXiv:2401.03468*, 2023.

Appendix A. Audio and Visual Utterance Pre-processing

For visual utterances, we only extract the lip region for AVSR. Following previous methods (Shi et al., 2022b; Afouras et al., 2018b,c), we use dlib (King, 2009) to detect 68 facial keypoints and align each face with its neighbors. We crop a 96×96 region of interest (ROI) centered on the mouth from each visual utterance. For audio utterances, we maintain the same pre-processing steps as in previous works (Ma et al., 2021a; Shi et al., 2022a). We extract 26-dimensional log filterbank energy features from the raw waveform and stack 4 neighboring acoustic frames for synchronization. During training, to enhance the data, we randomly crop an 88×88 region from the whole ROI and horizontally flip it with a probability of 0.5.

Pre-training Setup. AMG-AVSR is based on the pre-training process of AV-HuBERT (Shi et al., 2022b), directly utilizing its checkpoints for subsequent stages. During pre-training, we use a modified ResNet-18 (Ma et al., 2021a; Martinez, 2021) and a linear projection layer as visual and audio encoders, respectively. It considers two model configurations: Transformer-BASE and Transformer-LARGE, with 12/24 Transformer layers, embedding dimensions/feed-forward dimensions/attention heads of 768/3072/12 and 1024/4096/16, respectively. We simply adopted the pre-trained models obtained by training on LRS3 (Afouras et al., 2018c) and VoxCeleb2 (Chung et al., 2018).

Appendix B. Finetuning With Curriculum Learning Setup

To integrate two different curriculum learning strategies, we utilize multi-stage fine-tuning and observe their combined effects. Our experiments are conducted on 4 NVIDIA 4090 GPUs. First, we perform C-Modal curriculum learning. In this process, we freeze the encoder and train the decoder, with the masked information replaced by zero vectors. We fine-tune the audio modality alone for 20K/40K steps in the 29h/224h settings, respectively. Afterward, we unfreeze the encoder and continue C-Modal curriculum learning, training for an additional 20K/40K steps. Next, we gradually introduce noise to achieve C-Noise, training for 20K/40K steps until training is fully completed. For comparison, we adopt the same decoder configuration as (Shi et al., 2022b), utilizing a Transformer with 9 Transformer-decoder layers. Each stage is trained using Adam, with the learning rate being gradually warmed up to 0.0005 for the half of updates.