HPCNet: Hybrid Pixel and Contour Network for Audio-Visual Speech Enhancement With Low-Quality Video

Hang Chen, *Member, IEEE*, Chen-Yue Zhang, Qing Wang, *Member, IEEE*, Jun Du, *Senior Member, IEEE*, Sabato Marco Siniscalchi, *Senior Member, IEEE*, Shi-Fu Xiong, and Gen-Shun Wan

***Abstract*—To advance audio-visual speech enhancement (AVSE) research in low-quality video settings, we introduce the multimodal information-based speech processing-low quality video (MISP-LQV) benchmark, which includes a 120-hour real-world Mandarin audio-visual dataset, two video degradation simulation methods, and benchmark results from several well-known AVSE models. We also propose a novel hybrid pixel and contour network (HPCNet), incorporating a lip reconstruction and distillation (LRD) module and a contour graph convolution (CGConv) layer. Specifically, the LRD module reconstructs high-quality lip frames from low-quality audio-visual data, utilizing knowledge distillation from a teacher model trained on high-quality data. The CGConv layer employs spatio-temporal and semantic-contextual graphs to capture complex relationships among lip landmark points. Extensive experiments on the MISP-LQV benchmark reveal the performance degradation caused by low-quality video across various AVSE models. Notably, including real/simulated low-quality videos in AVSE training enhances its robustness to low-quality videos but degrades the performance of high-quality videos.The proposed HPCNet demonstrates strong robustness against video quality degradation, which can be attributed to (1) the reconstructed lip frames closely aligning with high-quality frames and (2) the contour features exhibiting consistency across different video quality levels. The generalizability of HPCNet also has been validated through experiments on the 2nd COG-MHEAR AVSE Challenge dataset.**

***Index Terms*—Speech enhancement, audio-visual, graph convolutional network, talking face generation, knowledge distillation.**

I. Introduction

THE human perceptual system can isolate a singular voice source even within highly noisy environments [1]. For instance, amidst the clamor of a restaurant, individuals are adept at concentrating on the discourse of their dining companion; during a fiery presidential debate, the ability to disentangle overlapping speech is evident. This impressive capability is rooted in the audio-visual nature of human language perception and production [2], [3]. Specifically, the human perceptual system heavily relies on visual cues to lessen noise in corrupted speech [4] and to direct attention towards an active speaker in a bustling environment [5]. The automation of this audio-visual speech enhancement (AVSE) holds substantial potential, with applications spanning from assistive technologies for the hearing impaired to enhancing auditory capabilities in wearable augmented reality devices and better transcription of spoken content in real-world videos.

Traditional AVSE methods can be traced back to the pioneering work [6], which initially demonstrated the advantages of incorporating visual features. Subsequently, more sophisticated frameworks based on classical statistical approaches were proposed, including [7], [8]. However, traditional AVSE algorithms often encounter challenges in effectively tracking unexpected nonstationary noise in real-world conditions. In recent years, data-driven AVSE approaches [9], [10] have increasingly leveraged deep neural networks (DNNs) [11], significantly outperforming traditional statistical methods. Specifically, DNN-based AVSE models build upon state-of-the-art (SOTA) audio-only speech enhancement (AOSE) models, optimizing visual feature selection and refining audio-visual fusion strategies. Common visual features include raw pixel values from the lip region of interest (ROI) [12], active appearance models [13], lip landmark points [14], [15], [16], and embeddings from the middle layer of a pre-trained word-level lipreading model [17], [18], [19], [20]. As for audio-visual fusion, concatenating audio and visual representations into a shared hidden layer has been a commonly used approach for AVSE. Other fusion strategies include addition-based fusion [21], product-based fusion [22], [23], squeeze-excitation fusion [24], [25], and attention-based fusion [22], [26].

However, existing research on AVSE assumes that videos are recorded in high quality. Nevertheless, video quality degradation is a crucial and common problem in real-world scenarios [27], [28], including resolution compression due to network latency or hardware limitations and errors in lip tracking due to occlusion or side-face views. Consequently, to advance research

Received 12 November 2024; revised 30 December 2024; accepted 28 March 2025. Date of publication 10 April 2025; date of current version 25 June 2025. This work was supported by the National Natural Science Foundation of China under Grant 62171427. The guest editor coordinating the review of this article and approving it for publication was Dr. Yu Tsao. *(Corresponding author: Jun Du.)*

Hang Chen, Chen-Yue Zhang, Qing Wang, and Jun Du are with the National Engineering Research Center of Speech and Language Information Processing, University of Science and Technology of China, Hefei 230027, China (e-mail: hangchen@ustc.edu.cn; cyzhang0722@mail.ustc.edu.cn; qingwang2@ustc.edu.cn; jundu@ustc.edu.cn).

Sabato Marco Siniscalchi is with the University of Palermo, 90133 Palermo, Italy (e-mail: sabatomarco.siniscalchi@unipa.it).

Shi-Fu Xiong and Gen-Shun Wan are with iFlytek Research, Hefei 230088, China (e-mail: sfxiong@iflytek.com; gswan@iflytek.com).

Dataset and source code will be available at https://github.com/coal-boss/HPCNet.

Digital Object Identifier 10.1109/JSTSP.2025.3559763

in AVSE with a focus on the robustness against low-quality video, this study first introduces a benchmark named multimodal information-based speech processing-low-quality video (MISP-LQV), featuring a Mandarin audio-visual dataset, two video degradation simulation methods, and benchmark results from several AVSE models. The dataset comprises 24 hours of audio-visual recordings from 263 native speakers in 26 homes, including clean speech alongside paired high- and low-quality video. Clean speech was mixed with real home noise at 5 different signal-to-noise ratios (SNRs), resulting in 120 hours of noisy data. Two simulation methods target missing and resolution degradation of the lip ROI. Performance degradation due to low-quality video is evident across different AVSE models. Additionally, we propose a novel hybrid pixel and contour network (HPCNet) that demonstrates strong robustness to low-quality video. Specifically, the lip reconstruction and distillation (LRD) module enhances video quality by utilizing noisy speech with additional knowledge distillation from a high-quality teacher model. The contour graph convolution (CGConv) layer utilizes spatio-temporal and semantic-contextual graphs to capture complex relationships among lip landmark points, which serve as contour features that complement traditional pixel-based visual features. Our main contributions can be summarized as follows:

1) releasing the MISP-LQV benchmark, which includes a 120-hour real-world Mandarin audio-visual dataset, two low-quality video simulation methods, and the results of several well-known AVSE models on this dataset.
2) proposing a novel hybrid pixel and contour network (HPCNet) that features a lip reconstruction and distillation (LRD) module to enhance video quality by utilizing noisy speech alongside the contour graph convolution (CGConv) layers to capture the complex relationships among lip landmark points.
3) confirming the challenges posed by low-quality video and the effectiveness of HPCNet through a series of experiments conducted on the MISP-LQV benchmark, along with the generalizability of HPCNet through additional experiments on the 2nd COG-MHEAR AVSE Challenge dataset.

The remainder of the paper is organized as follows. Section II reviews related works. Section III describes the released MISP-LQV benchmark. Section IV presents our proposed HPCNet. Section V analyses the experimental results. Finally, we summarize our findings in Section VI.

II. Related Work

Since its inception, AVSE has made significant progress, with early works [29], [30], [31], [32], [33], [34], [35] laying the foundation. In recent years, data-driven AVSE approaches have increasingly leveraged deep neural networks (DNNs) to process high-dimensional visual data [11]. Most DNN-based AVSE models focus on visual feature selection and audio-visual fusion strategies, building on SOTA AOSE models.

In terms of visual features, early DNN-based AVSE models [12], [36] effectively harness raw pixel values from the target speaker's lip region of interest (ROI), allowing for practical visual information extraction through end-to-end training. Despite achieving notable advancements over audio-only baselines, these models demand extensive amounts of paired audio-visual data for training [37], [38] and face optimization challenges due to the high-dimensional input [25]. Consequently, various dimensionality reduction techniques have been introduced, including the active appearance model (AAM) [39] and the 2D discrete cosine transform (2D-DCT) [40]. However, dimensionality reduction leads to loss of details, potentially hindering AVSE performance. Lip landmark points [14] have been employed as low-dimensional visual features due to their sparse nature. Furthermore, the motion of these landmark points [15], [16] has proven beneficial. Nevertheless, traditional DNNs, recurrent neural networks (RNNs), and convolutional neural networks (CNNs) consistently struggle to effectively model the spatial relationships between these discrete landmark points, highlighting an area for improvement.

Currently, various pre-trained models are widely adopted to extract low-dimensional features from the raw pixel space. These pre-training methods can be categorized into supervised pre-training and self-supervised learning. Supervised pre-training includes tasks such as face recognition [9], word-level lipreading [17], phone-level lipreading [18], audio-visual articulation place recognition [19], and audio-visual phone and articulation place recognition [20]. On the other hand, self-supervised learning approaches include models for learning audio-visual correlation evidence [41], audio-visual temporal synchronization [42], [43], deep multi-instance, multi-label learning [44], and audio-visual object segmentation [45].

For audio-visual fusion strategies, while certain studies have demonstrated DNN's capacity to perform both early fusion [15] and late fusion [13], [36], the most common approach is to integrate audio and visual representations into a shared hidden layer, known as intermediate fusion [46]. Specifically, concatenation has been a popular method [10], [21], [47], [48], [49], [50]. Other fusion strategies include addition-based fusion [13], [21], product-based fusion [22], [23], [51], and squeeze-excitation fusion [24], [25]. Attention-based fusion has also been investigated to select the more informative modality [26] flexibly, including additive attention [22], temporal attention [48], [52], spatial-wise attention [52], factorized attention [53], and rule-based attention [53].

III. MISP-LQV Benchmark

The introduced MISP-LOV benchmark is curated explicitly for home scenarios and consists of paired high- and low-quality video recordings captured by two cameras positioned at varying distances. In addition, this research proposes two methods for simulating low-quality video based on real-world data analysis: missing frames and resolution degradation. A detailed description of the design process, statistical information, and simulation methods is presented herein.

A. Dataset Design

As shown in Fig. 1, a group of people is engaged in conversation while watching television in a cozy living room. Meanwhile,

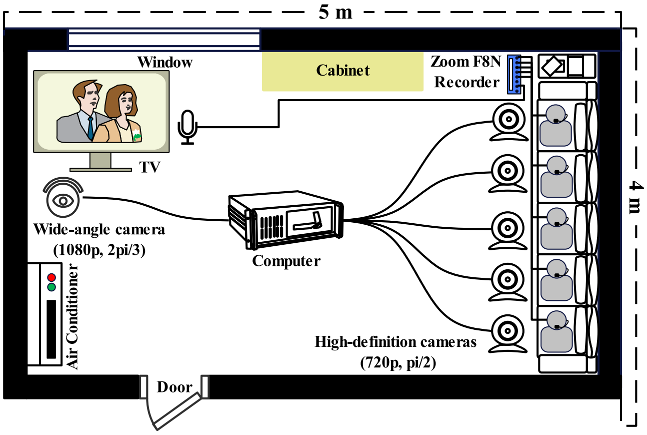

5 m
4 m
Window
Cabinet
Zoom F8N Recorder
TV
Wide-angle camera (1080p, 2pi/3)
Computer
Air Conditioner
High-definition cameras (720p, pi/2)
Door

Fig. 1. An example of the recording venue and the used devices.

an AVSE system strives to enhance the clarity and quality of each speaker's speech before transmitting it to communication channels or intelligent assistants. However, capturing video in real home settings presents numerous challenges, including potential blurriness, furniture obstructions, and low lighting conditions. These factors often result in diminished video quality, making it an ideal scenario for evaluating the resilience of AVSE systems in handling low-quality video. In the context of home settings, we constructed the MISP-LQV dataset using two cameras positioned at varying distances to capture paired video recordings of both high and low quality. Furthermore, we employed two sets of microphones to separately record clean speech and background noise, followed by traditional noisy speech simulation techniques to generate precisely aligned noisy-clean speech pairs for SE model training and evaluation.

Specifically, a wide-angle camera was situated 3 to 5 meters away from the speaker for low-quality recording, with a field of view of 141° diagonally, 120° horizontally, and 63° vertically. It recorded video at a resolution of 1920 × 1080 pixels at a rate of 25 frames per second, encompassing the entire indoor environment. However, this configuration resulted in a low-resolution lip region of interest (ROI) frequently obstructed by surrounding objects. Conversely, a high-definition camera was positioned 0.5 to 0.8 meters in front of the speaker to capture high-quality video, with a field of view of 116° diagonally, 99° horizontally, and 53.4° vertically, recorded at 1280 × 720 resolution and 25 frames per second. The near-field camera focused solely on the speaker, ensuring a visible and high-resolution lip ROI. All cameras were synchronized by connecting them to a central computer utilizing Vicando software. Each speaker was equipped with a high-fidelity directional microphone attached under the chin, recording at a sampling rate of 44.1 kHz and a bit depth of 16 bits. The clean speech was seldom interfered with by off-target sources and had a signal-to-noise ratio (SNR) exceeding 25 dB. Each speaker used a high-fidelity directional microphone under the chin, recording at a 44.1 kHz sampling rate and 16-bit depth. This setup ensured minimal interference from other sources, achieving an SNR greater than 25 dB. An omnidirectional microphone near the television on the opposite side of the room also captured ambient noise. Both microphones were connected to a ZOOM F8n sound card for clock synchronization. Manual alignment was employed to synchronize audio and video: at the start of each session, a distinct synchronization cue was produced by tapping a cup, and during post-processing, the frame showing the contact was aligned with the corresponding impact sound in the waveform.

TABLE I
Details of Clean Data in MISP-LQV Benchmark

Dataset	**Train-clean**	**Dev-clean**	**Eval-clean**	**Total**
Duration (h)	20.27	1.86	1.77	23.9
Utterance	70, 728	5, 742	5, 730	82200
Room	18	2	6	26
Participant	233	15	15	263
-Male	72	6	8	86
-Female	105	15	18	138

Upon completing the recording process, we compiled approximately 100 hours of raw audio-visual data and transcribed the accompanying text. Data cleaning was performed utilizing DNSMOS P.835 [54], an objective metric to evaluate speech quality in wideband scenarios. The long-duration recordings were segmented and segments exhibiting an overall quality score exceeding 4.2 were preserved, yielding 27 hours of high-quality speech. A secondary manual inspection was conducted to ensure the absence of discernible noise artifacts, resulting in 24 hours of clean speech data. These clean speech recordings and their corresponding video were subsequently categorized into train-clean, dev-clean, and eval-clean sets with no overlap in speakers or rooms. Detailed statistics for each subset are summarized in Table I. Then, we mixed 4 hours of background noise with 70,215 utterances from the train-clean set at signal-to-noise ratios (SNRs) of 10, 5, 0, -5 and -10 dB. The noise came from 18 rooms in the train-clean set, resulting in a 101.35-hour training set. We applied the same method for the development and evaluation sets. In the dev-clean set, we mixed 6,255 utterances with 0.8 hours of noise from the same rooms, creating a 9.3-hour development set. In the eval-clean set, we mixed 5,730 utterances with 1.1 hours of noise, leading to an 8.85-hour evaluation set.

B. Low-Quality Video Simulation

Fig. 2 serves as an example illustrating a comparison between a pair of high- and low-quality videos. We can observe two primary degradation patterns:

- *Missing Lip ROI:* The lip ROI could not be detected due to factors such as lighting conditions, changes in pose, and transmission channel issues.
- *Lip ROI Resolution Degradation:* The resolution of the lip ROIs was reduced due to factors such as camera quality and shooting distance.

Based on these observations, we developed two low-quality video simulators to augment the availability of training data. Fig. 2 illustrates the simulated lip frames that have undergone frame missing and low resolution simulations. A visual evaluation indicates that no distinguishable difference between real and simulated low-quality lip frames.

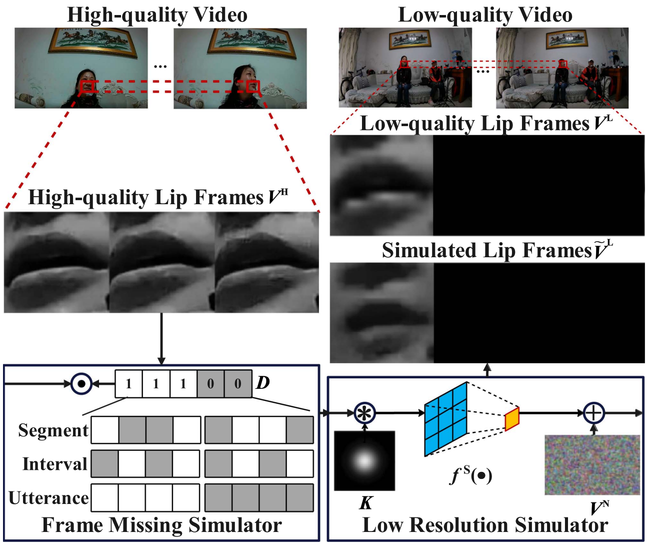

High-quality Video
Low-quality Video
Low-quality Lip Frames V^{L}
High-quality Lip Frames V^{H}
Simulated Lip Frames $\widetilde{V}^{\mathrm{L}}$
1 1 1 0 0 D
Segment
Interval
Utterance
Frame Missing Simulator
K
$f^{\mathrm{S}}(\bullet)$
V^{N}
Low Resolution Simulator

Fig. 2. An example comparing high-quality, low-quality, and simulated lip frames. Combining two low-quality video simulation methods involving lip ROI missing and resolution degradation has resulted in a high degree of visual similarity between the simulated and real low-quality lip frames.

In the frame missing simulator, given a sequence of high-quality lip frames $V^{\mathrm{H}} = [v_1^{\mathrm{H}}, \ldots, v_t^{\mathrm{H}}, \ldots, v_{T_{\mathrm{V}}}^{\mathrm{H}}]$, covering T_{V} frames, each $v_t \in \mathbb{R}^{H\times W}$ represents a gray-scale lip image. The simulator first generates a mask sequence $D = [d_1, \ldots, d_t, \ldots, d_T]$, where $d_t \in \{0, 1\}$: "0" indicates a dropped frame and "1" indicates a retained frame. The simulated lip frames $\widetilde{V}^{\mathrm{L}} \in \mathbb{R}^{T_{\mathrm{V}}\times H\times W}$ is calculated as follows:

$$\widetilde{V}^{\mathrm{L}} = V^{\mathrm{H}} \odot D \tag{1}$$

where $\odot$ denotes the Hadamard product. There are three ways to generate D:

- *Segment:* For each v_t^{H}, a missing probability α_t is randomly sampled from a uniform distribution. The corresponding d_t is calculated as follows:

$$d_t = \begin{cases} 1, u_t > \beta \\ 0, u_t \leq \beta \end{cases} \tag{2}$$

where $\beta \in [0, 1]$ presents the preset frame missing rate.
- *Utterance:* For each sample V^{H}, a missing probability α is randomly sampled from a uniform distribution. The corresponding D is calculated as follows:

$$D = \begin{cases} [1, \ldots, 1, \ldots, 1], u > \alpha \\ [0, \ldots, 0, \ldots, 0], u \leq \alpha \end{cases} \tag{3}$$

- *Interval:* For each sample V^{H}, a missing probability α_t is randomly sampled from a uniform distribution. $d_t \in D$, it is calculated as follows:

$$d_t = \begin{cases} 1, \text{other} \\ 0, \alpha_t \leq \beta \text{ and } t \bmod \lceil 1/\beta \rceil = 0 \end{cases} \tag{4}$$

where mod and $\lceil\cdot\rceil$ denote the modulo operation and the ceiling function, respectively.

By enriching the patterns and frequencies of lip frame omissions, our simulation method effectively covers real-world missing scenarios, including lip occlusion, changes in camera angles, and transmission/memory issues.

In the low resolution simulation, each v_t^{H} undergoes blurring, downsampling, and noise injection sequentially. The entire process can be formulated as follows:

$$\tilde{v}_t^{\mathrm{L}} = f^{\mathrm{S}}(v_t^{\mathrm{H}} * K) + V^{\mathrm{N}} \tag{5}$$

Where K represents a square Gaussian blur kernel, the size of the kernel directly impacts the level of blurring, with larger sizes resulting in more pronounced blurring effects. Additionally, $f^{\mathrm{S}}(\cdot)$ denotes a bicubic downsampling function, where the downsampling factor dictates the extent of pixel reduction. A higher downsampling factor leads to a more substantial decrease in pixels. Furthermore, V^{N} refers to visual noise, encompassing Gaussian or salt-and-pepper noise, and its variance determines the impact on image details. A higher variance obscures more image details.

IV. Hybrid Pixel and Contour Network for AVSE

This section presents an in-depth overview of the proposed HPCNet designed for AVSE. We will first highlight the holistic framework of HPCNet, showcasing its network architecture and training process. Then, we will dive into two pivotal components of HPCNet: the detailed structure and pre-training methodology of the LRD module along with the structural specifics of the CGConv layer.

A. Overall Framework

As shown in Fig. 3(a), given the low-quality gray-scale lip frames $V^{\mathrm{L}} \in \mathbb{R}^{T_{\mathrm{V}}\times H\times W}$ and the noisy FBANK feature $F^{\mathrm{N}} \in \mathbb{R}^{T_{\mathrm{A}}\times C_{\mathrm{M}}}$, the pre-trained LRD module utilizes an audio-visual embedding extractor (AVEE) followed by a reconstructor to produce reconstructed grayscale lip frames $V^{\mathrm{R}} \in \mathbb{R}^{T_{\mathrm{V}}\times H\times W}$:

$$V^{R} = \mathrm{Reconstructor}\left(\mathrm{AVEE}\left(V^{\mathrm{L}}, F^{\mathrm{N}}\right)\right) \tag{6}$$

where T_{A} and C_{M} represent the audio sequence length and the number of mel filters, respectively. In this study, $C_{\mathrm{M}} = 40$ are set as default. V^{R} are subsequently fed back into the LRD module and combined with F^{N} to yield the refined audio-pixel embedding $E^{\mathrm{R}} \in \mathbb{R}^{T_{\mathrm{A}}\times C_{\mathrm{E}}}$ via the internal AVEE module:

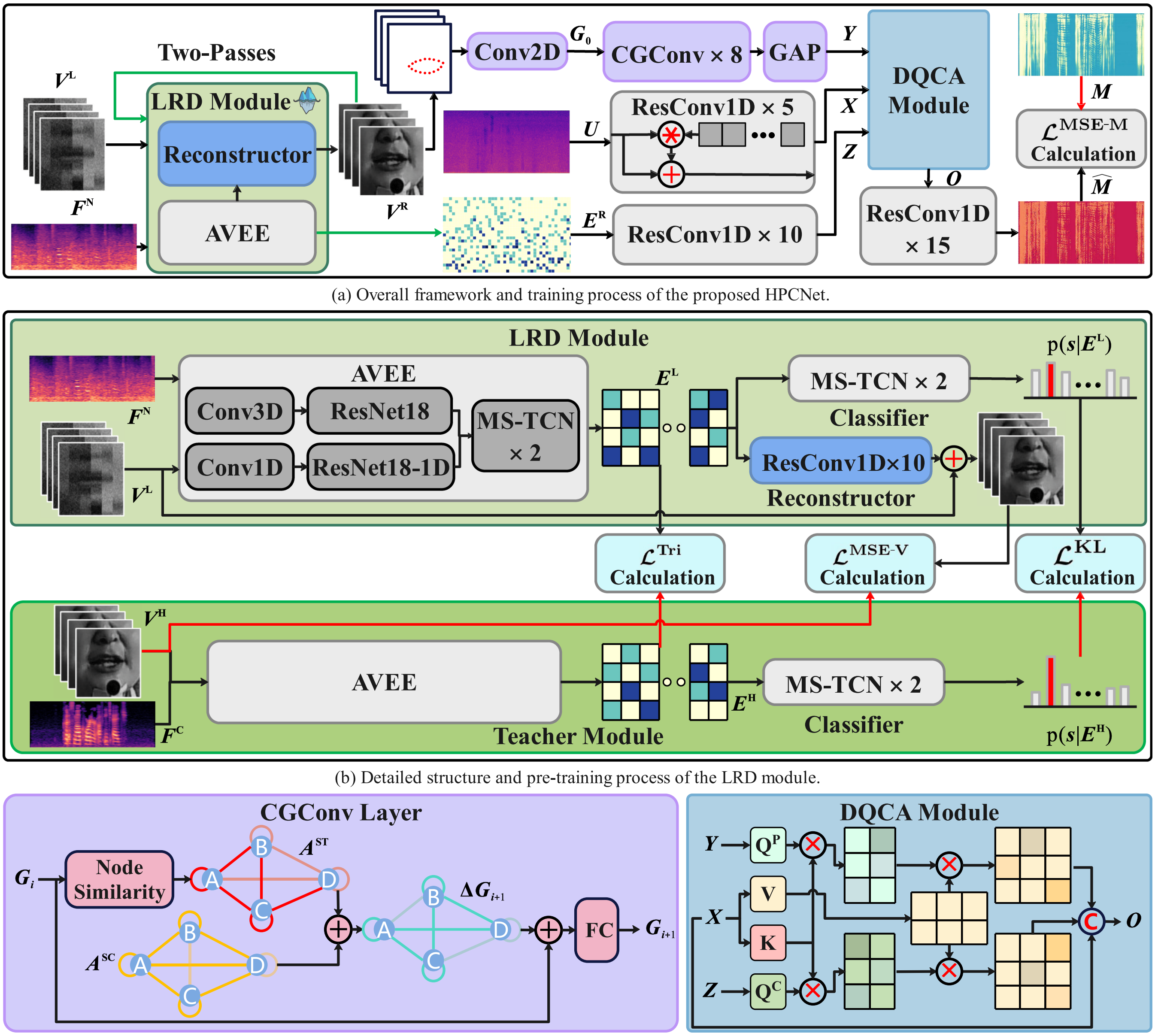
$$E^{\mathrm{R}} = \mathrm{AVEE}\left(V^{\mathrm{R}}, F^{\mathrm{N}}\right) \tag{7}$$

where C_{E} denotes the channel number of the audio-pixel embedding, set to 512 by default. Then, E^{R} is sent to 10 stacked ResConv1D blocks for learning the high-level audio-pixel representation $Z \in \mathbb{R}^{T_{\mathrm{A}}\times C_{\mathrm{E}}}$ as follows:

$$Z = \mathrm{ResConv1D}_{10}\left(\cdots \mathrm{ResConv1D}_{1}\left(E^{\mathrm{R}}\right)\right) \tag{8}$$

where each ResConv1D block consists of a 1D convolution layer with a residual connection, followed by a pReLU activation and batch normalization, as described in [19]. Additionally, 5 stacked ResConv1D blocks are employed to extract deep speech representation $X \in \mathbb{R}^{T_{\mathrm{A}}\times C_{\mathrm{E}}}$ from the noisy log power spectrum (LPS) feature $U \in \mathbb{R}^{T_{\mathrm{A}}\times C_{\mathrm{S}}}$ as follows:

$$X = \mathrm{ResConv1D}_{5}\left(\cdots \mathrm{ResConv1D}_{1}(U)\right) \tag{9}$$



(a) Overall framework and training process of the proposed HPCNet.

(b) Detailed structure and pre-training process of the LRD module.

(c) Detailed structure of the CGConv layer.

(d) Detailed structure of the DQCA module.

Fig. 3. Illustration of the proposed HPCNet for AVSE, including (a) The overall framework and the training process of the HPCNet, (b) Detailed structure and pre-training process of the LRD module, (c) Detailed structure of the CGConv layer and (d) Detailed structure of the DQCA module. V^{L}, V^{H} and V^{R}: high-quality, low-quality and reconstructed gray-scale lip frames. F^{N} and F^{C}: noisy and clean FBANK features. E^{H}, E^{L} and E^{R}: high-quality, low-quality and refined audio-pixel embeddings. G_i: i-layer contour feature. U: noisy LPS feature. X, Y, Z and O: deep speech, contour, audio-pixel and fused representations. $\hat{M}$ and M: predicted magnitude and ideal ratio masks.

where C_{S} represents the number of frequency bins in the spectrogram, set to 257 by default in this study.

To incorporate lip contour, we select N_{L} landmark points in the lip ROI and the pixel coordinates of each point are input to a 2D convolution layer to extract the initial contour feature $G_0 \in \mathbb{R}^{T_{\mathrm{V}} \times K \times C_{\mathrm{E}}}$:

$$G_0 = \mathrm{Conv2D}\left(\mathrm{LandmarkPoints}\left(V^{\mathrm{R}}\right)\right) \tag{10}$$

where N_{L} is set to 20 by default. The kernel size and stride of the 2D convolution are $(5, 7)$ and $(1, 1)$, respectively. Traditional CNNs often fall short of capturing the intricate relationships between discrete points. However, graph convolutional networks (GCNs) excel in modeling these complex interactions. Leveraging this advantage, we propose a novel contour graph convolutional (CGConv) layer to extract meaningful insights from lip contour features. In the HPCNet, eight stacked CGConv layers followed by a global average pooling (GAP) layer transform the initial contour feature G_0 into a deep contour representation $Y \in \mathbb{R}^{T_{\mathrm{V}} \times C_{\mathrm{E}}}$:

$$Y = \mathrm{GAP}\left(\mathrm{CGConv}_8\left(\cdots \mathrm{CGConv}_1\left(G_0\right)\right)\right) \tag{11}$$

Next, a novel dual-query cross-attention (DQCA) module is proposed to fuse the intermediate representations generated by different branches, namely the contour representation Y, the audio-pixel representation Z, and the speech representation X. Specifically, Y and Z serve as query inputs, whereas X acts as both the key and value inputs. The detailed fusion process is carried out as follows:

$$O = \text{Repeat-Concat}\left(D_1^{\text{P}}, D_1^{\text{C}}, \cdots, D_{N_{\text{A}}}^{\text{P}}, D_{N_{\text{A}}}^{\text{C}}, X\right)\mathbf{W}^{\text{O}} \quad (12)$$

$$D_n^{\text{P}}, D_n^{\text{C}} = W_n^{\text{P}} X \mathbf{W}_n^{\text{V}}, W_n^{\text{C}} X \mathbf{W}_n^{\text{V}} \quad (13)$$

$$W_n^{\text{P}} = \text{SoftMax}\left(\frac{Z\mathbf{Q}_n^{\text{P}}\left(X\mathbf{W}_n^{\text{K}}\right)^{\top}}{\sqrt{C_{\text{A}}}}\right) \quad (14)$$

$$W_n^{\text{C}} = \text{SoftMax}\left(\frac{Y\mathbf{Q}_n^{\text{C}}\left(X\mathbf{W}_n^{\text{K}}\right)^{\top}}{\sqrt{C_{\text{A}}}}\right) \quad (15)$$

where $O \in \mathbb{R}^{T_{\text{A}} \times C_{\text{E}}}$ and $\mathbf{W}^{\text{O}} \in \mathbb{R}^{(2N_{\text{A}}C_{\text{A}}+C_{\text{E}}) \times C_{\text{E}}}$ denote the fused representation and the output projection matrix, respectively. N_{A} attention heads are used, and $n \in \{1, \ldots, N_{\text{A}}\}$ is the index of the head. For the n-th attention head, $D_n^{\text{P}} \in \mathbb{R}^{T_{\text{A}} \times C_A}$ and $D_n^{\text{C}} \in \mathbb{R}^{T_{\text{V}} \times C_{\text{A}}}$ denote the output feature maps. $W_n^{\text{P}} \in \mathbb{R}^{T_{\text{A}} \times T_{\text{A}}}$ and $W_n^{\text{C}} \in \mathbb{R}^{T_{\text{V}} \times T_{\text{A}}}$ denote the attention weight. $Q_n^{\text{P}} \in \mathbb{R}^{C_{\text{E}} \times C_A}$, $Q_n^{\text{C}} \in \mathbb{R}^{C_{\text{E}} \times C_A}$, $\mathbf{W}_n^{\text{K}} \in \mathbb{R}^{C_{\text{E}} \times C_A}$ and $\mathbf{W}_n^{\text{V}} \in \mathbb{R}^{C_{\text{E}} \times C_A}$ are the projection matrices of query, key and value. C_{A} is the number of channel in the DQCA module. The mismatch in sequence length between D_n^{P} and D_n^{C}, i.e., $T_A \neq T_V$, is resolved by repeating each frame of D_n^{C} across multiple frames of D_n^{P}.

Finally, O is passed through a stack of 15 ResConv1D blocks to predict a magnitude mask $\hat{M} \in \mathbb{R}^{T_{\text{A}} \times C_{\text{M}}}$:

$$\hat{M} = \text{Sigmoid}\left(\text{ResConv1D}_{15}\left(\cdots \text{ResConv1D}_1(O)\right)\right) \quad (16)$$

The predicted mask $\hat{M}$ can be used to filter the noisy spectrum and reconstruct the waveform by inverse short-time Fourier transform (iSTFT). The MSE between the ideal ratio mask (IRM) [55] $M \in \mathbb{R}^{T_{\text{A}} \times C_{\text{M}}}$ and $\hat{M}$ is used as the loss function, denoted as MSE-M, which is computed as follows:

$$\mathcal{L}^{\text{MSE-M}} = \frac{1}{T_{\text{A}} C_{\text{M}}} \sum_{t=1}^{T_{\text{A}}} \sum_{j=1}^{C_{\text{M}}} \left|\hat{m}_{t,j} - m_{t,j}\right|^2 \quad (17)$$

where $\hat{m}_{t,j}$ and $m_{t,j}$ are the values at the t-th frame and j-th frequency bin of $\hat{M}$ and M, respectively. Note that the LRD model does not update its parameters during training with MSE-M. In subsequent experiments, we also constructed two models using only contour features as input: the contour graph network (CGNet) and the contour convolution cetwork (CCNet). Specifically, CGNet differs from HPCNet by excluding the LRD module, the audio-pixel embedding E^R, and the following 10 ResConv1D blocks, while CCNet replaces the 8 CGConv layers in CGNet with 8 ResConv1D blocks.

B. Lip Reconstruction and Distillation

As shown in Fig. 3(b), the LRD module comprises an AVEE module, a classifier, and a reconstructor. The AVEE module encompasses visual, audio, and fusion branches, utilizing a dual-tower architecture as outlined in [19]. The visual branch incorporates a spatiotemporal convolution, followed by an 18-layer ResNet. Similarly, the audio branch adopts a comparable structure, with 1D kernels replacing the 3D kernels in the spatiotemporal convolution and ResNet18 utilizing 1D kernels instead of 2D kernels. The outputs from the visual and audio branches are concatenated and channeled into a 2-layer multi-scale temporal convolutional network (MS-TCN), acting as the fusion branch. Architectural details have been omitted in our description, but more information can be found in [19].

And the classifier consists of a 2-layer MS-TCN [56] followed by a softmax activation function and predicts the posterior probability of the articulation place $p(\boldsymbol{s}|E^{\text{L}}) \in \mathbb{R}^{T_{\text{A}} \times N_{\text{P}}}$ based the audio-pixel embedding $E^{\text{L}} \in \mathbb{R}^{T_{\text{A}} \times C_{\text{E}}}$, which is extracted with the low-quality lip frames and noisy speech data. The prediction process can be described as follows:

$$p(\boldsymbol{s}|E^{\text{L}}) = \text{SoftMax}\left(\text{MS-TCN}_2\left(\text{MS-TCN}_1\left(E^{\text{L}}\right)\right)\right) \quad (18)$$

where $\boldsymbol{s} = [s_1, \ldots, s_{T_A}]$ represents the articulation places sequence, and N_{P} denotes the size of the dictionary.

We also integrated a reconstructor parallel to the classifier with 10 stacked ResConv1D blocks to predict improved grayscale lip frames $V^{\text{R}} \in \mathbb{R}^{T_{\text{V}} \times H \times W}$. Furthermore, we incorporated a residual connection from the original low-quality lip frames V^{L} to ease the training process:

$$V^{\text{R}} = V^{\text{L}} + \text{ResConv1D}_{10}\left(\cdots \text{ResConv1D}_1\left(E^{\text{L}}\right)\right) \quad (19)$$

MSE measure the distortion of V^{R} compared to the high-quality lip frames V^{H}, referred to as MSE-V:

$$\mathcal{L}^{\text{MSE-V}} = \frac{1}{T_{\text{V}} H L} \sum_{\tau=1}^{T_{\text{V}}} \sum_{h=1}^{H} \sum_{h=l}^{H} \left|v_{t,h,l}^{\text{R}} - v_{t,h,l}^{\text{H}}\right|^2 \quad (20)$$

where $v_{t,h,l}^{\text{R}}$ and $v_{t,h,l}^{\text{H}}$ are the values at the t-th frame, h-th row and l-th column of V^{R} and V^{H}.

We further introduce an intermediate distillation framework. As illustrated in Fig. 3(b), the teacher model also features an AVEE module followed by a classifier and takes high-quality lip frames V^{H} and clean FBANK feature $F^{\text{C}} \in \mathbb{R}^{T_{\text{A}} \times C_{\text{M}}}$ as inputs, resulting in high-quality audio-pixel embeddings $E^{\text{H}} \in \mathbb{R}^{T_{\text{A}} \times C_{\text{E}}}$ and an accurate posterior probability estimation of the articulation place $p(\boldsymbol{s}|E^{\text{H}}) \in \mathbb{R}^{T_{\text{A}} \times N_{\text{P}}}$. The entire teacher model is trained to minimize cross-entropy loss:

$$\mathcal{L}^{\text{CE}} = -\sum_{t=1}^{T_{\text{A}}} \log p(s_t|\boldsymbol{e}_t^{\text{H}}) \quad (21)$$

where $\boldsymbol{e}_t^{\text{H}} \in \mathbb{R}^{N_{\text{P}}}$ is the t-th feature vector of E^{H}.

In the distillation process, E^{H} and $p(\boldsymbol{s}|E^{\text{H}})$ serve as carriers of valuable knowledge. The knowledge transfer is quantified using a triplet loss $\mathcal{L}^{\text{Tri}}$ between E^{H} and E^{L}, as well as a Kullback-Leibler (KL) divergence loss $\mathcal{L}^{\text{KL}}$ between $p(\boldsymbol{s}|E^{\text{L}})$ and $p(\boldsymbol{s}|E^{\text{H}})$. Specifically, for $\forall \boldsymbol{e}_t^{\text{L}} \in E^{\text{L}}$, the temporally corresponding feature vector $\boldsymbol{e}_t^{\text{H}}$ in E^{H} are treated as positive samples, while others are considered negative. The triplet loss is computed as follows:

$$\mathcal{L}^{\mathrm{Tri}} = \max\left(0, \sum_{t=1}^{T_{\mathrm{A}}}\left(\sum_{\upsilon=1}^{C_{\mathrm{E}}}(e^{\mathrm{L}}_{t,\upsilon} - e^{\mathrm{H}}_{t,\upsilon})^2 - \frac{1}{T_{\mathrm{A}}-1}\sum_{\varphi=1\wedge\varphi\neq t}^{T_{\mathrm{A}}}\sum_{\upsilon=1}^{C_{\mathrm{E}}}(e^{\mathrm{L}}_{t,\upsilon} - e^{\mathrm{H}}_{\varphi,\upsilon})^2 + \alpha\right)\right) \tag{22}$$

where $e^{\mathrm{L}}_{t,\upsilon}$ and $e^{\mathrm{H}}_{t,\upsilon}$ denote the values at the υ-th channel bin of $\boldsymbol{e}^{\mathrm{L}}_t$ and $\boldsymbol{e}^{\mathrm{H}}_t$, respectively. α represents the margin, which controls the gap between the distance to positive and negative samples. Additionally, $\mathcal{L}^{\mathrm{KL}}$ can be described as follows:

$$\mathcal{L}^{\mathrm{KL}} = \sum_{t=1}^{T_{\mathrm{A}}} p(s_t|\boldsymbol{y}^{\mathrm{L}}_t)\log\frac{p(s_t|\boldsymbol{e}^{\mathrm{L}}_t)}{p(s_t|\boldsymbol{e}^{\mathrm{H}}_t)} \tag{23}$$

Finally, the total loss function is defined as a combination of MSE-V, triplet loss, and KL divergence, calculated as follows:

$$\mathcal{L}^{\mathrm{Tol}} = \mathcal{L}^{\mathrm{MSE-V}} + \mathcal{L}^{\mathrm{Tri}} + \mathcal{L}^{\mathrm{KL}} \tag{24}$$

By minimizing $\mathcal{L}^{\mathrm{Tol}}$, the reconstructed lip frames steadily approach high-quality standards.

C. CGConv Layer

We explore the intricate relationships between lip landmark points from two distinct perspectives. The first perspective delves into the spatial-temporal relationships among landmark points, which rely heavily on the current articulated content. Consequently, we define spatio-temporal graph adjacency matrices as sample-dependent graph adjacency matrices. The other perspective focuses on the semantic contextual relationships, contingent on the statistical findings of semantic context from the training data. As a result, we define articulation context graph adjacency matrices as sample-independent graph adjacency matrices. Based on the above analysis, the CGConv layer is ingeniously designed to incorporate a spatio-temporal graph (STG) and a semantic-contextual graph (SCG) to capture their respective relationships effectively. Both graph structures are meticulously parameterized and can be optimized in conjunction with other network parameters end-to-end. Furthermore, the CGConv layer features a fully connected layer and a residual connection, enhancing feature representation and facilitating the training of the entire network. Consequently, we can articulate the forward process of the i-th CGConv layer as follows:

$$G_i = (G_{i-1} + \Delta G_i)\,\mathbf{W}^{\mathrm{FFN}}_i \tag{25}$$

where $G_{i-1} \in \mathbb{R}^{T_{\mathrm{V}}\times K\times C_{\mathrm{E}}}$ and $G_i \in \mathbb{R}^{T_{\mathrm{V}}\times K\times C_{\mathrm{E}}}$ represent the input and output feature maps, respectively. $\mathbf{W}^{\mathrm{FFN}}_i \in \mathbb{R}^{C_{\mathrm{E}}\times C_{\mathrm{E}}}$ denotes the parameters of the fully connected layer. $\Delta G_i \in \mathbb{R}^{T_{\mathrm{V}}\times K\times C_{\mathrm{E}}}$ represents the output of the graph convolution on the input feature G_{i-1} and can be formulated as:

$$\Delta G_i = \Lambda_i^{\mathrm{ST}^{-\frac{1}{2}}}\left(A_i^{\mathrm{ST}} + \mathbf{I}\right)\Lambda_i^{\mathrm{ST}^{-\frac{1}{2}}} G_{i-1}\mathbf{W}^{\Delta}_i + \mathbf{\Lambda}_i^{\mathrm{SC}^{-\frac{1}{2}}}\left(\mathbf{A}_i^{\mathrm{SC}} + \mathbf{I}\right)\mathbf{\Lambda}_i^{\mathrm{SC}^{-\frac{1}{2}}} G_{i-1}\mathbf{W}^{\Delta}_i \tag{26}$$

where $A_i^{\mathrm{ST}} \in \mathbb{R}^{T_{\mathrm{V}}\times K\times K}$ and $\mathbf{A}_i^{\mathrm{SC}} \in \mathbb{R}^{K\times K}$ represent the adjacency matrix of the SCG and STG, respectively. As previously described, A_i^{ST} is sample-dependent, meaning each frame has its dedicated adjacency matrix. In contrast, $\mathbf{A}_i^{\mathrm{SC}}$ is sample-independent, with all frames sharing the same adjacency matrix. $\mathbf{I}$ is the identity matrix with the same dimensions as the adjacency matrix and $\mathbf{W}^{\Delta}_i \in \mathbb{R}^{C_{\mathrm{E}}\times C_{\mathrm{E}}}$ represents the feature transformation matrix. $\Lambda_i^{\mathrm{ST}} \in \mathbb{R}^{T_{\mathrm{V}}\times K\times K}$ and $\mathbf{\Lambda}_i^{\mathrm{SC}} \in \mathbb{R}^{K\times K}$are degree matrices of $A_i^{\mathrm{ST}} + \mathbf{I}$ and $\mathbf{A}_i^{\mathrm{SC}} + \mathbf{I}$, respectively, which are used to normalize the adjacency matrices. The normalized adjacency matrices $\bar{A}_i^{\mathrm{ST}}$ and $\bar{\mathbf{A}}_i^{\mathrm{SC}}$ can be expressed as follows:

$$\bar{A}_i^{\mathrm{ST}} = \Lambda_i^{\mathrm{ST}^{-\frac{1}{2}}}\left(A_i^{\mathrm{ST}} + \mathbf{I}\right)\Lambda_i^{\mathrm{ST}^{-\frac{1}{2}}}$$
$$\bar{\mathbf{A}}_i^{\mathrm{SC}} = \mathbf{\Lambda}_i^{\mathrm{SC}^{-\frac{1}{2}}}\left(\mathbf{A}_i^{\mathrm{SC}} + \mathbf{I}\right)\mathbf{\Lambda}_i^{\mathrm{SC}^{-\frac{1}{2}}} \tag{27}$$

Accordingly, (26) can be rewritten as:

$$\Delta G_i = (\bar{A}_i^{\mathrm{ST}} + \bar{\mathbf{A}}_i^{\mathrm{SC}})G_{i-1}\mathbf{W}^{\Delta}_i \tag{28}$$

From (28), CGConv calculation can be divided into two steps: node features are transformed by a learnable parameter matrix $\mathbf{W}^{\Delta}_i$ and nodes features are aggregated by a specific normalized adjacency matrix $\bar{A}_i^{\mathrm{ST}} + \bar{\mathbf{A}}_i^{\mathrm{SC}}$.

STG captures the changing shape of speakers' lips over time. The normalized adjacency matrix $\bar{A}_i^{\mathrm{ST}}$ is created by comparing the similarity between pairs of nodes. To do this, we use a soft attention mechanism to calculate node spatial similarity. The formula for computing $\bar{A}_i^{\mathrm{ST}}$ is as follows:

$$\bar{A}_i^{\mathrm{ST}} = \mathrm{SoftMax}\left[\left(G_{i-1}\mathbf{W}^{\theta}_i\right)\left(G_{i-1}\mathbf{W}^{\phi}_i\right)\right]^{\top} \tag{29}$$

where $\mathbf{W}^{\theta}_i, \mathbf{W}^{\phi}_i \in \mathbb{R}^{C_{\mathrm{E}}\times C_{\mathrm{G}}}$ are the parameters of the embedding spaces θ and ϕ, respectively. C_{G} represents the dimension of the embedding space, set to 256 by default.

Regarding the semantic contextual relationship is inherent of the talking mouth, we assume that the SCG is a fully connected graph, and all parameters of the normalized adjacency matrix $\bar{\mathbf{A}}_i^{\mathrm{SC}}$ are learned from the training data without any prior assumption. Therefore, we initialize $\bar{\mathbf{A}}_i^{\mathrm{SC}}$ with a constant value, as shown below:

$$\bar{\mathbf{A}}_i^{\mathrm{SC}} = c\mathbf{J} \tag{30}$$

where $\mathbf{J}$ represents the all-ones matrix. c represents the initialized constant, set to 0.000001 by default.

V. Experiments and Results Analysis

To evaluate speech quality and intelligibility, we employed PESQ and STOI, respectively. PESQ [57] applies an auditory transform to compare the loudness spectrum of clean and enhanced speech, yielding a score ranging from -0.5 to 4.5; higher scores indicate better speech quality. The STOI [58] compares the temporal envelopes of clean and enhanced speech in short-time regions, providing values between 0 and 1, with higher values representing better speech intelligibility.

For the training strategy, we used the Adam [59] optimizer for 100 epochs, implementing early stopping if there was no improvement in the development loss for 10 consecutive epochs.

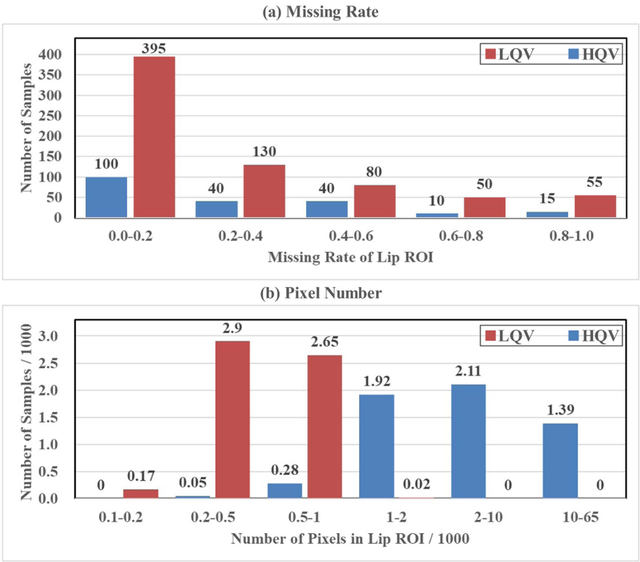


Fig. 4. Comparison of the missing rate (upper panel) and pixel count (lower panel) of the lip ROI between paired high- and low-quality videos in the MISP-LQV evaluation set. **HQV**: high-quality video, **LQV**: low-quality video.

The initial learning rate was set to 0.0003 and halved during training if there was no improvement for 3 epochs in the development loss. Regarding training data, we adopted a $1:1$ ratio of real high-quality and low-quality videos in each training batch to balance the model's performance. Additionally, we employed an online simulation strategy for generating simulated data, where each training batch randomly selects one simulation method and configuration.

A. Performance Analysis of Data

1) Quantitative Analysis of Video Quality Degradation: We initially performed comprehensive quantitative analyses to identify the differences between high- and low-quality videos. Specifically, through the results obtained from lip ROI detection and tracking, we derived the proportion of frames in each video where the lip ROI was undetectable, as well as the average pixel count within the lip ROI. The bar charts in Fig. 4 compare these two metrics between paired high- and low-quality videos in the MISP-LQV evaluation set. Based on Fig. 4(a), it's clear that most high-quality video frames successfully detect the lip ROI, with only 205 samples exhibiting missing lip ROIs, accounting for just 3.5% of the evaluation set. In contrast, low-quality videos show 395 problematic samples with a missing rate in the range of $(0, 0.2]$, nearly double the number of problematic samples in high-quality videos. The total number of problematic samples in low-quality videos is 710, representing 12.4% of the evaluation set. Among the most severe 55 problematic samples, over 80% of video frames failed to detect a valid lip ROI, leading to a complete loss of crucial visual semantic information for AVSE. Meanwhile, Fig 4(b) presents a more pronounced gap between high- and low-quality videos in terms of lip ROI pixel count. In high-quality videos, the lip ROI is typically represented by 2,000 to 10,000pixels, whereas in low-quality videos, the pixel count drops to around $200 - 500$. This reduction in pixel count leads to the blurring of lip movement details, which may degrade performance in AVSE.

TABLE II
COMPARISON OF PESQ AND STOI (IN %) AMONG NOISY, NOEASE, AVPL-VIR, MTMEASE AND MEASES WITH DIFFERENT TRAINING VIDEO ON THE MISP-LQV EVALUATION SET

Model	**Training Video**			**PESQ ↑**		**STOI (%) ↑**	
	HQV	**LQV**	**SIM**	**HQV**	**LQV**	**HQV**	**LQV**
Noisy		\		1.68		69.51	
NoEASE		\		2.44		83.44	
AVPL-VIR [60]	✓	×	×	2.67	2.27	86.68	82.41
MTMEASE [20]	✓	×	×	2.70	2.29	86.63	82.30
MEASE [19]	✓	×	×	2.71	2.32	86.71	82.54
	✓	✓	×	2.57	2.51	84.33	83.86
	✓	✓	✓	**2.63**	**2.58**	**85.84**	**84.29**

SIM: simulated video.

2) Performance Comparison Between High- and Low-Quality Videos: We assessed the performance of various AVSE models in both high- and low-quality video conditions. Table II presents the comparison of PESQ and STOI (in %) for the unprocessed system denoted as Noisy, NoEASE, AVPL-VIR, MTMEASE, and MEASE on the MISP-LQV evaluation set. It is evident that several well-known AVSE models, including AVPL-VIR, MTMEASE, and MEASE, experienced significant performance degradation when evaluated with low-quality video, which resulted in PESQ declines of 0.40, 0.41, and 0.39, respectively, while their STOI scores dropped by 4.27%, 4.33% and 4.17%, respectively. The decline in performance demonstrates the considerable influence of video quality on AVSE effectiveness, highlighting the critical importance of improving AVSE systems' robustness to low-quality video. Given that MEASE achieved the best overall performance, the subsequent experiments will primarily focus on it.

To gain deeper insights into the factors affecting AVSE performance in low-quality videos, we categorized all samples in the low-quality evaluation set into 5 groups based on their missing rate of the lip ROI and 6 groups based on the number of pixels in the lip ROI. Fig. 5 presents a comparison of PESQ and STOI (in %) across different missing rate and pixel number groups enhanced by MEASE. Our observations reveal a clear downward trend in both PESQ and STOI as the lip ROI missing rate increases and the pixel count decreases, underscoring the detrimental impact of lip ROI absence and reduced pixel count on AVSE performance.

3) Analysis of Training With Low-Quality Videos: Further, we assessed the impact of incorporating low-quality video during training, including real and simulated samples. Table II also provides PESQ and STOI scores (in %) for MEASEs trained with real and simulated low-quality videos on the MISP-LQV evaluation set. Our findings revealed that integrating real low-quality video into training improved the robustness of MEASE under low-quality video evaluation, which resulted in absolute improvements of 0.19 in PESQ and 1.32% in STOI, respectively.

However, mixed-quality training presented a new challenge that the quality and intelligibility of the enhanced speech

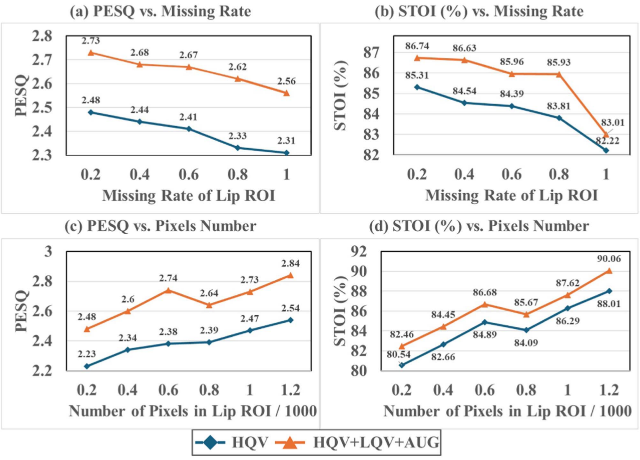


Fig. 5. Comparison of PESQ and STOI (in %) among different missing rate and pixels number groups on the on the MISP-LQV low-quality evaluation set enhanced by MEASEs with different training video: (a) PESQ vs. missing rate, (b) STOI (%) vs. missing rate, (c) PESQ vs. pixels number and (d) STOI (%) vs. pixels number.

TABLE III
Comparison of PESQ and STOI (In %) Between MEASEs and HPCNets With LRD Ablations on the MISP-LQV Evaluation Set

Model	LRD		PESQ ↑		STOI (%) ↑	
	TPE	IKD	HQV	LQV	HQV	LQV
MEASE-Large	×	×	2.62	2.56	85.68	84.17
MEASE	×	×	2.63	2.58	85.84	84.29
	✓	×	2.66	2.62	86.24	85.91
	✓	✓	2.67	2.66	86.45	86.27
HPCNet	×	×	2.73	2.69	87.01	86.68
	✓	×	2.75	2.72	87.26	87.00
	✓	✓	**2.76**	**2.74**	**87.41**	**87.21**

TPE: two-pass extraction, IKD: intermediate knowledge distillation.

deteriorated under high-quality video, with decreases of 0.14 in PESQ and 2.38% in STOI. This decline could be attributed to the shared labels between paired high- and low-quality videos, which caused an average output across different-quality video inputs. Moreover, incorporating simulated low-quality video consistently improved the performance of MEASE in both high-quality and low-quality evaluations, yielding absolute gains of 0.06 and 0.07 in PESQ, and 1.51% and 0.43% in STOI, respectively. Further results in Fig. 5 reveal that these improvements were consistently observed across different levels of video degradation. These findings highlight varying degrees of degradation help bridge the gap between real high- and low-quality videos, contributing to a balance between robustness and performance under varying video quality conditions.

B. Performance Analysis of HPCNet

1) Overall Comparison: We first explore the performance of the proposed HPCNet and present a comparison of PESQ and STOI (in %) results for MEASE and HPCNet in Table III. It is evident that HPCNet significantly outperforms MEASE across various video quality levels. Specifically, in evaluations with high-quality video, HPCNet achieves absolute gains of 0.13 in PESQ and 1.57% in STOI compared to MEASE. Furthermore, in evaluations with low-quality video, the improvements facilitated by HPCNet are even more pronounced, with enhancements of 0.16 in PESQ and 2.92% in STOI, respectively. These findings underscore the superior robustness of HPCNet against low-quality video compared to MEASE. Furthermore, considering that HPCNet involves an increase in parameters compared to MEASE, we also developed a large version of the MEASE model by proportionally increasing the number of ResConv1D blocks by 62%. However, MEASE-Large performed worse than MEASE, with PESQ and STOI reductions of 0.01 and 0.16% under high-quality video, and 0.02 and 0.12% under low-quality video, respectively. This degradation suggests that simply increasing the parameter count may not improve performance. By contrast, thoughtful architectural design, namely HPCNet with the LRD module and CGConv layer, delivers substantial performance gains compared to MEASE-Large at a similar parameter scale. Specifically, HPCNet achieves PESQ and STOI improvements of 0.14 and 1.73% under high-quality video and 0.18 and 3.04% under low-quality video, respectively. In the following sections, we will conduct ablation experiments to analyze the contributions of two key components in the HPCNet: the LRD module and the CGConv layer.

2) Ablation Study on LRD Module: Table III also presents a comparison of PESQ and STOI (in %) results for MEASEs and HPCNets with and without the LRD module. The findings unequivocally demonstrate that the LRD module significantly improves performance across various models and video quality levels. Notably, in the evaluation with high-quality video, the MEASE model achieved absolute gains of 0.04 in PESQ and 0.61% in STOI, after incorporating the LRD module. In comparison, the ablation of the LRD module resulted in a noticeable performance decline for HPCNet, with a loss of 0.03 in PESQ and a 0.4% decrease in STOI, respectively. Moreover, in the evaluation with low-quality video, the LRD module yielded even more pronounced improvements, which narrowed the performance gap between high-quality and low-quality evaluations. When added to MEASE, it shows enhancements of 0.08 in PESQ and 1.98% in STOI, whereas its removal from HPCNet results in declines of 0.05 in PESQ and 0.53% in STOI. These findings underscore the contribution of the LRD module in fortifying the robustness of AVSE models to low-quality video, demonstrating its generalizability across different models. Additionally, higher SNR speech cues enable the reconstruction of higher-quality lip frames, further improving AVSE performance. More experimental results can be found in our released code repository.

Furthermore, we analyzed the specific benefits of the LRD module under low-quality video conditions by conducting ablation experiments on its key components, including the two-pass extraction (TPE) and intermediate knowledge distillation (IKD). The results in Table III show the PESQ and STOI (in %) scores for the MEASE and HPCNet when these key modules are ablated. It was observed that under low-quality video evaluation conditions, the TPE with the reconstructed lip frames significantly improved the performance of both MEASE and HPCNet.

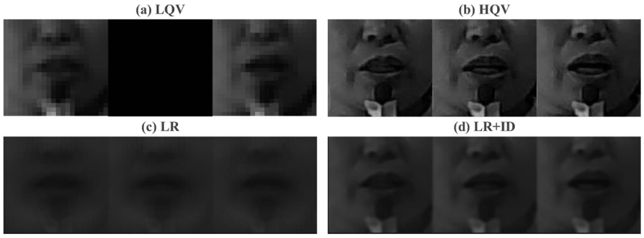


Fig. 6. A example comparing low-quality, reconstructed, and high-quality lip frame sequences: (a) LQV, (b) HQV, (c) LRD w./o. IKD and (d) LRD.

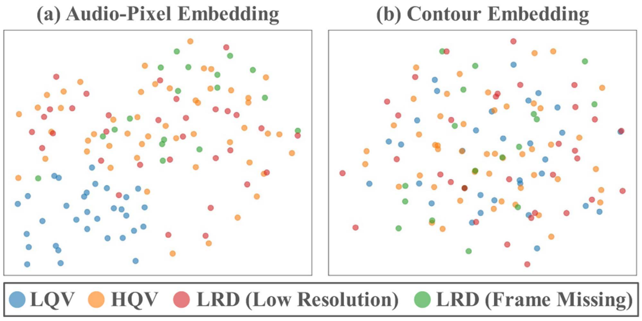


Fig. 7. Comparative t-SNE analysis of audio-pixel and contour embeddings for low-quality, reconstructed and high-quality lip frames: (a) audio-pixel embedding and (b) contour embedding.

Specifically, MEASE showed increases of 0.04 in PESQ and 1.62% in STOI, while HPCNet exhibited absolute gains of 0.03 in PESQ and 0.32% in STOI. These results highlight the reconstructed lip frames provided an enriched visual context, which can effectively bridge the quality gap under challenging conditions. Moreover, the application of IKD led to further improvements of 0.04 in PESQ and 0.36% in STOI for MEASE, and 0.05 in PESQ and 0.53% in STOI for HPCNet. We speculate that intermediate distillation helps to generate more precise reconstructed video frames.

Fig. 6 shows a comparison of low-quality, reconstructed, and high-quality lip frame sequences. The reconstructed lip frames successfully restore a complete lip structure. Moreover, through the integration of intermediate distillation, the reconstructed frames exhibit markedly crisper lip details, demonstrating a substantial enhancement in visual fidelity.

We also analyzed the differences between these lip sequences as they passed through HPCNet. Fig. 7 presents a comparison of audio-pixel and contour embeddings among the low-quality, reconstructed, and high-quality lip frames. Note that the audio-pixel and contour embeddings corresponding to the missing lip frames in the low-quality video are omitted for clarity. It is evident that the difference in audio-pixel embeddings between the raw low-quality and high-quality frames is striking. However, the reconstructed audio-pixel embeddings closely mirror the high-quality ones, demonstrating the effective restoration of details lost due to video quality degradation, such as missing frames and reduced resolution. The negligible differences between low-resolution and high-resolution contour embeddings validate our previous findings about lip landmarks. This confirms the robustness of contour features to resolution degradation. Significantly, the reconstructed contour embeddings closely mirror the original high-quality ones, affirming that the reconstructed lip frame sequences align with high-quality sequences both visually and in the information extracted by the model. These findings underscore the effectiveness of the proposed LRD module.

TABLE IV
Comparison of PESQ and STOI (In %) Among MEASE, CCNet, CGNet and HPCNets With CGConv Ablations on the MISP-LQV Evaluation Set

Model	APE	Contour Feature			PESQ ↑		STOI (%) ↑	
		CNN	ASG	STG	HQV	LQV	HQV	LQV
MEASE	✓	×	×	×	2.63	2.58	85.84	84.29
CCNet	×	✓	×	×	2.54	2.50	83.91	83.62
CGNet	×	×	✓	✓	2.60	2.57	85.27	84.03
HPCNet	✓	×	✓	×	2.71	2.69	87.01	86.17
	✓	×	×	✓	2.74	2.71	87.33	86.45
	✓	×	✓	✓	**2.76**	**2.74**	**87.41**	**87.21**

APE: audio-pixel embedding, ASG: adaptive semantic graph, STG: spatio -temporal graph.

3) Ablation Study on CGConv Layer: We first conducted ablation experiments on the key components of the CGConv layer, including ASG and STG. Table IV also shows the PESQ and STOI (in %) results for HPCNets when the ASG or STG was removed. It is evident that removing the ASG led to degradations across different video qualities, with absolute reductions of 0.02 in PESQ and 0.08% in STOI under high-quality video evaluation and reductions of 0.03 in PESQ and 0.76% in STOI under low-quality conditions. Conversely, ablating the STG resulted in even more significant declines, with absolute reductions of 0.05 in PESQ and 0.40% in STOI during evaluations of high-quality video. For low-quality conditions, the reductions were 0.05 in PESQ and 1.04% in STOI. These findings emphasize the effectiveness of the CGConv layer in extracting valuable information from contour features, indicating that the spatiotemporal relationships are more crucial than the semantic relationships.

To better understand the advantages of contour features in HPCNet, we also conducted a comprehensive comparison of the average PESQ and STOI (in %) scores between the MEASE and CGNet, which exclusively utilize pixel and contour features as inputs, respectively. Regrettably, as illustrated in Table IV, CGNet declined compared to MEASE across all video quality conditions. Specifically, PESQ and STOI decreased by 0.03 and 0.57% under high-quality video, respectively. However, this decline was less pronounced under low-quality video conditions, with reductions of 0.01 in PESQ and 0.26% in STOI. These findings lead us to speculate that contour features may offer significant robustness only against specific video degradation scenarios. Furthermore, Table IV compares different modeling approaches using the same contour feature input, specifically CGNet using CGConv layers versus CCNet using ResConv1D blocks. We can observe that CGNet outperforms CCNet, where PESQ and STOI improved by 0.06 and 1.36% under high-quality video and by 0.13 and 0.41% under low-quality video. This

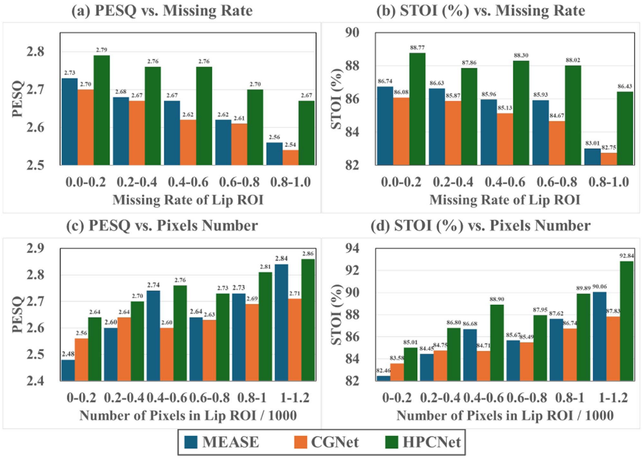


Fig. 8. Comparison of PESQ and STOI (in %) among different missing rate and pixels number groups on the on the MISP-LQV low-quality evaluation set enhanced by MEASE, CGNet and HPCNet: (a) PESQ vs. missing rate, (b) STOI (%) vs. missing rate, (c) PESQ vs. pixels number and (d) STOI (%) vs. pixels number.

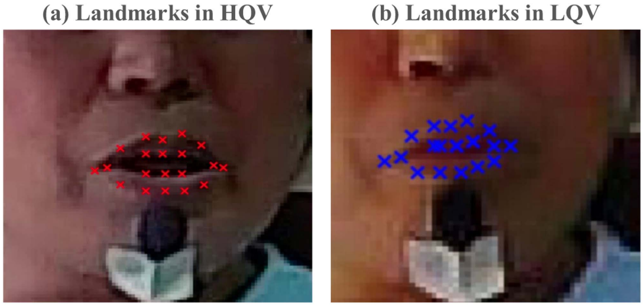


Fig. 9. A frame example comparing lip landmark distribution across videos with different quality: (a) Landmarks in HQV and (b) Landmarks in LQV.

TABLE V
COMPARISON OF PESQ AND STOI (IN %) AMONG ALL SUBMISSIONS, MEASE AND HPCNET ON THE 2ND COG-MHEAR AVSE CHALLENGE EVALUATION SET

Model	PESQ ↑	STOI (%) ↑
Noisy	1.14	44.10
Baseline	1.41	55.63
Enhanced_AVSE	1.15	44.34
ENU_JHU	1.20	46.56
WLV	1.27	49.82
Rezzsl	1.28	50.28
BioASP_CITI	1.41	53.78
Diffusion	1.45	54.03
ICT_AVSU	1.66	67.67
AVSE02	1.61	68.21
MERL	**2.71**	83.76
MEASE	2.37	81.21
HPCNet	2.68	**84.27**

improvement demonstrates the superiority of GCNs over traditional CNNs for modeling complex relationships among lip landmarks.

To verify this hypothesis, we divided all samples in the low-quality evaluation set into 5 groups based on the missing rate of the lip ROI and 6 groups based on the number of pixels in the lip ROI. Fig. 8 compares average PESQ and STOI (in %) across different missing rate and pixel number groups, which were enhanced by the MEASE, CGNet, and HPCNet. We observed that CGNet showed greater robustness to low-resolution degradation than MEASE. Specifically, in the $200 - 400$ pixel range, CGNet outperformed MEASE with improvements of 0.04 in PESQ and 0.30% in STOI, respectively. In the $0 - 200$ pixel range, CGNet exhibited even more significant improvements with gains of 0.08 in PESQ and 1.13% in STOI, respectively. In contrast, for pixel number ranges above 400, the MEASE surpassed CGNet, attaining higher PESQ and STOI scores. These findings highlight the complementary relationship between contour and pixel features across varying resolution conditions. HPCNet effectively leverages this complementary nature, achieving the highest PESQ and STOI scores across all resolution levels.

The observations can be attributed to the inherent characteristics of contour and pixel features. The contour feature relies on spatial relationships among lip landmarks and offers sparse yet consistent descriptions even at reduced resolutions. Fig. 9 illustrates an example of paired high- and low-resolution samples, showing that the lip landmark distribution remains nearly unchanged across different resolutions. In contrast, the pixel feature excels at capturing detailed information with its dense nature, providing richer cues as resolution increases.

4) Generalizability Study: To evaluate the generalizability of the proposed HPCNet across different datasets, we extended our assessment to include the 2nd COG-MHEAR AVSE Challenge [61] dataset. The challenge provides approximately 113 hours of mixed speech and video for the training set and 8.5 hours for the development set. Audio tracks of interferers are composed of a single competing speaker or a noise source in the following ranges: -15 dB to 5 dB (competing speaker) and -10 dB to 10 dB (noise). The videos of the target speakers and the competing speakers in the training set are selected from the LRS3 dataset [62]. Noise data mainly comes from Clarity Challenge (1st) [63], Freesound [64], and DNS Challenge (2nd) [65]. There are 650 target speakers, 405 competing speakers and 7,346 noise files in the training set. The development set has 85 target speakers, 30 competing speakers and 1825 noise files. All audio files are monaural speech with a 16 kHz sampling frequency and 16-bit depth.

Table V compares PESQ and STOI (in %) among all submissions, MEASE, and HPCNet on the official evaluation set. The evaluation set has 1,389 extracted sentences from 30 speakers (15 females and 15 males). Approximately half of the mixed speech in the evaluation set has a competing speaker scenario while the other half has noise. There are six competing speakers (3 females and 3 males). The noise types used in the evaluation set are a subset of the noise types used in the training and development sets. The results clearly show that HPCNet outperformed the MEASE baseline, yielding gains of 0.31 in PESQ and 3.06% in STOI, respectively. Notably, HPCNet achieved the highest STOI score and the second-highest PESQ score compared to other submissions. Specifically, compared to the MERL team, HPCNet had a 0.51 higher STOI but a 0.04 lower PESQ. It's important to highlight that HPCNet still maintained a significant advantage in the PESQ score compared to all submissions except MERL. This consistent trend of superiority underscores the outstanding effectiveness and adaptability of HPCNet across diverse evaluation benchmarks. Additionally, upon manually inspecting the videos in the evaluation set, we found that most videos were high-quality, featuring low frame-missing and high resolutions. Consequently, the pixel features captured more detailed lip movements, explaining why HPCNet did not achieve even more significant improvements in this scenario.

VI. Conclusion

This study advances AVSE research for real-world low-quality video scenarios by releasing the MISP-LQV benchmark and proposing HPCNet. The MISP-LQV benchmark comprises 120 hours of paired high- and low-quality Mandarin audio-visual recordings from 263 speakers across 26 homes, along with two video quality degradation simulators: frame missing and resolution reduction. Experimental findings demonstrate that (1) several well-known AVSE models exhibit consistent performance degradation under low-quality video conditions, and (2) training with real and simulated low-quality videos improves robustness against low-quality videos but degrades performance for high-quality videos. HPCNet exhibits strong robustness to video quality degradation, thanks to the LRD module and CGConv layer. The LRD module enhances video quality by leveraging noisy speech with additional knowledge distillation from a high-quality teacher model. The CGConv layer captures complex relationships among lip landmark points using spatio-temporal and semantic-contextual graphs, remaining reliable at low resolutions. In future work, we plan to explore integrating pretrained foundation models into our audio-visual feature extraction pipeline to capitalize on their rich representational capacity. While this approach promises further performance gains, it substantially increases parameter counts and computational overhead. We will investigate strategies such as model compression and knowledge distillation to maintain a practical balance between enhanced performance and resource efficiency.

References

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *J. Acoustical Soc. Amer.*, vol. 25, no. 5, pp. 975–979, 1953.

[2] D. G. Stork and M. E. Hennecke, *Speechreading by Humans and Machines: Models, Systems, and Applications*, vol. 150. Berlin, Germany: Springer Science & Business Media, 2013.

[3] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[4] T. Rahne, M. Böckmann, H. Von, and E. S. Sussman, "Visual cues can modulate integration and segregation of objects in auditory scene analysis," *Brain Res.*, vol. 1144, pp. 127–135, 2007.

[5] E. Z. Golumbic, G. B. Cogan, C. E. Schroeder, and D. Poeppel, "Visual input enhances selective speech envelope tracking in auditory cortex at a cocktail party," *J. Neurosci.*, vol. 33, no. 4, pp. 1417–1426, 2013.

[6] T. Darrell, J. W. Fisher, and P. Viola, "Audio-visual segmentation and "the cocktail party effect"," in *Proc. Int. Conf. Multimodal Interfaces*, 2000, pp. 32–40.

[7] B. Rivet, L. Girin, and C. Jutten, "Mixing audiovisual speech processing and blind source separation for the extraction of speech signals from convolutive mixtures," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 96–108, Jan. 2007.

[8] B. Rivet, L. Girin, and C. Jutten, "Visual voice activity detection as a help for speech source separation from convolutive mixtures," *Speech Commun.*, vol. 49, no. 7–8, pp. 667–677, 2007.

[9] A. Ephrat et al., "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 1–11, 2018.

[10] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," in *Proc. Interspeech*, 2018, pp. 3244–3248.

[11] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[12] A. Gabbay, A. Shamir, and S. Peleg, "Visual speech enhancement," in *Proc. Interspeech* 2018, pp. 1170–1174.

[13] F. U. Khan, B. P. Milner, and T. Le Cornu, "Using visual speech information in masking methods for audio speaker separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1742–1754, Oct. 2018.

[14] J.-C. Hou et al., "Audio-visual speech enhancement using deep neural networks," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2016, pp. 1–6.

[15] G. Morrone, S. Bergamaschi, L. Pasa, L. Fadiga, V. Tikhanoff, and L. Badino, "Face landmark-based speaker-independent audio-visual speech enhancement in multi-talker environments," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6900–6904.

[16] Y. Li, Z. Liu, Y. Na, Z. Wang, B. Tian, and Q. Fu, "A visual-pilot deep fusion for target speech separation in multitalker noisy environment," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. 2020*, 2020, pp. 4442–4446.

[17] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," in *Proc. Interspeech* 2018, pp. 3244–3248.

[18] J. Wu et al., "Time domain audio visual speech separation," in *Proc. Autom. Speech Recognit. Understanding Workshop* 2019, pp. 667–673.

[19] H. Chen et al., "Correlating subword articulation with lip shapes for embedding aware audio-visual speech enhancement," *Neural Netw.*, vol. 143, pp. 171–182, 2021.

[20] C. Wang, H. Chen, J. Du, B. Yin, and J. Pan, "Multi-task joint learning for embedding aware audio-visual speech enhancement," in *Proc. ISCSLP 2022*, 2022, pp. 255–259.

[21] T. Afouras, J. S. Chung, and A. Zisserman, "My lips are concealed: Audio-visual speech enhancement through obstructions," in *Proc. Interspeech*, 2019, pp. 4295–4299.

[22] T. Ochiai, M. Delcroix, K. Kinoshita, A. Ogawa, and T. Nakatani, "Multimodal SpeakerBeam: Single channel target speech extraction with audio-visual speaker clues," in *Proc. Interspeech*, 2019, pp. 2718–2722.

[23] W. Wang, C. Xing, D. Wang, X. Chen, and F. Sun, "A robust audio-visual speech enhancement model," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. 2020*, 2020, pp. 7529–7533.

[24] H. R. V. Joze, A. Shaban, M. L. Iuzzolino, and K. Koishida, "MMTM: Multimodal transfer module for CNN fusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. 2020*, 2020, pp. 13289–13299.

[25] M. L. Iuzzolino and K. Koishida, "AV(SE)^2: Audio-visual squeeze-excite speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. 2020*, 2020, pp. 7539–7543.

[26] C. Li and Y. Qian, "Deep audio-visual speech separation with attention mechanism," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. 2020*, 2020, pp. 7314–7318.

[27] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 2017, pp. 6447–6456.

[28] J. Hong, M. Kim, J. Choi, and Y. M. Ro, "Watch or listen: Robust audio-visual speech recognition with visual corruption modeling and reliability scoring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. 2023*, 2023, pp. 18783–18794.

[29] L. Girin, G. Feng, and J.-L. Schwartz, "Noisy speech enhancement with filters estimated from the speaker's lips," in *Proc. Eurospeech* 1995, pp. 1559–1562.

[30] L. Girin, J.-L. Schwartz, and G. Feng, "Audio-visual enhancement of speech in noise," *J. Acoustical Soc. Amer.*, vol. 109, no. 6, pp. 3007–3020, 2001.

[31] J. W. Fisher III, T. Darrell, W. Freeman, and P. Viola, "Learning joint statistical models for audio-visual fusion and segregation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, pp. 772–778.

[32] S. Deligne, G. Potamianos, and C. Neti, “Audio-visual speech enhancement with AVCDCN (audio-visual codebook dependent cepstral normalization),” in *Proc. Sensor Array Multichannel Signal Process. Workshop Proc.*, 2002, pp. 68–71.

[33] R. Goecke, G. Potamianos, and C. Neti, “Noisy audio feature enhancement using audio-visual speech data,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2002, vol. 2, pp. 2025–2028.

[34] J. Hershey and M. Casey, “Audio-visual sound separation via hidden Markov models,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, vol. 14, pp. 1173–1180.

[35] A. H. Abdelaziz, S. Zeiler, and D. Kolossa, “Twin-HMM-based audio-visual speech enhancement,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 3726–3730.

[36] A. Gabbay, A. Ephrat, T. Halperin, and S. Peleg, “Seeing through noise: Visually driven speaker separation and enhancement,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 3051–3055.

[37] B. İnan, M. Cernak, H. Grabner, H. P. Tukuljac, R. C. Pena, and B. Ricaud, “Evaluating audiovisual source separation in the context of video conferencing,” in *Proc. Interspeech*, 2019, pp. 4579–4583.

[38] Y. Luo, J. Wang, X. Wang, L. Wen, and L. Wan, “Audio-visual speech separation using i-vectors,” in *Proc. 2nd Int. Conf. Inf. Commun. Signal Process.* 2019, pp. 276–280.

[39] T. F. Cootes, G. J. Edwards, and C. J. Taylor, “Active appearance models,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.

[40] A. Adeel, M. Gogate, and A. Hussain, “Contextual deep learning-based audio-visual switching for speech enhancement in real-world environments,” *Inf. Fusion*, vol. 59, pp. 163–170, 2020.

[41] J.-W. Xiong, Y. Zhou, P. Zhang, L. Xie, W. Huang, and Y. Zha, “Look&listen: Multi-modal correlation learning for active speaker detection and speech enhancement,” *IEEE Trans. Multimedia*, vol. 25, pp. 5800–5812, 2023.

[42] Z. Pan, R. Tao, C. Xu, and H. Li, “Selective listening by synchronizing speech with lips,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 1650–1664, 2022.

[43] A. Owens and A. A. Efros, “Audio-visual scene analysis with self-supervised multisensory features,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 631–648.

[44] R. Gao, R. Feris, and K. Grauman, “Learning to separate object sounds by watching unlabeled video,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 35–53.

[45] T. Afouras, A. Owens, J. S. Chung, and A. Zisserman, “Self-supervised learning of audio-visual objects from video,” in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 208–224.

[46] D. Ramachandram and G. W. Taylor, “Deep multimodal learning: A survey on recent advances and trends,” *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 96–108, Nov. 2017.

[47] A. Arriandiaga et al., “Audio-visual target speaker extraction on multi-talker environment using event-driven cameras,” in *Proc. ISCAS*, 2021, pp. 1–5.

[48] S.-W. Chung, S. Choe, J. S. Chung, and H. G. Kang, “FaceFilter: Audio-visual speech separation using still images,” in *Proc. Interspeech* 2020, pp. 3481–3485.

[49] S.-Y. Chuang, Y. Tsao, C.-C. Lo, and H. M. Wang, “Lite audio-visual speech enhancement,” in *Proc. Interspeech* 2020, pp. 1131–1135.

[50] M. Gogate, A. Adeel, R. Marxer, J. Barker, and A. Hussain, “DNN driven speaker independent audio-visual mask estimation for speech separation,” in *Proc. Interspeech* 2018, pp. 2723–2727.

[51] R. Lu, Z. Duan, and C. Zhang, “Listen and look: Audio–visual matching assisted speech source separation,” *IEEE Signal Process. Lett.*, vol. 25, no. 9, pp. 1315–1319, Sep. 2018.

[52] Z. Sun, Y. Wang, and L. Cao, “An attention based speaker-independent audio-visual deep learning model for speech enhancement,” in *Proc. Int. Conf. Multimedia Model.*, 2020, pp. 722–728.

[53] R. Gu, S.-X. Zhang, Y. Xu, L. Chen, Y. Zou, and D. Yu, “Multi-modal multi-channel target speech separation,” *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 3, pp. 530–541, Mar. 2020.

[54] C. K. Reddy, V. Gopal, and R. Cutler, “Dnsmos p. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 886–890.

[55] C. Hummersone, T. Stokes, and T. Brookes, “On the ideal ratio mask as the goal of computational auditory scene analysis,” in *Blind Source Separation: Advances in Theory, Algorithms and Applications*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 349–368.

[56] B. Martinez, P. Ma, S. Petridis, and M. Pantic, “Lipreading using temporal convolutional networks,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6319–6323.

[57] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2001, vol. 2, pp. 749–752.

[58] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time–frequency weighted noisy speech,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[59] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–15.

[60] C.-Y. Zhang, H. Chen, J. Du, B.-C. Yin, J. Pan, and C.-H. Lee, “Incorporating visual information reconstruction into progressive learning for optimizing audio-visual speech enhancement,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.

[61] A. L. A. Blanco et al., “AVSE challenge: Audio-visual speech enhancement challenge,” in *Proc. Spoken Lang. Technol.*, 2023, pp. 465–471.

[62] T. Afouras, J. S. Chung, and A. Zisserman, “LRS3-TED: A large-scale dataset for visual speech recognition,” CoRR, abs/1809.00496, 2018.

[63] S. Graetzer et al., “Clarity-2021 challenges: Machine learning challenges for advancing hearing aid processing,” in *Proc. Interspeech*, 2021, pp. 686–690.

[64] J. Thiemann, N. Ito, and E. Vincent, “The diverse environments multichannel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings,” in *Proc. Meetings Acoust.*, 2013, vol. 19, no. 1, Art. no. 035081.

[65] C. K. A. Reddy et al., “ICASSP 2021 deep noise suppression challenge,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 6623–6627.

Hang Chen (Member, IEEE) received the B.Eng. and Ph.D. degrees from the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), Hefei, China, in 2018 and 2024, respectively. He is currently a Postdoctoral Researcher with USTC. His research focuses on audio-visual speech enhancement and recognition.

Chen-Yue Zhang received the B.Eng. degree in electronic information engineering from Xidian University, Xi’an, China, in 2018. She is currently working toward the master’s degree with the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, China. Her research focuses on audio-visual speech enhancement.

Qing Wang (Member, IEEE) received the B.S. and Ph.D. degrees from the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), Hefei, China, in 2012 and 2018, respectively. From 2018 to 2020, she worked on speech enhancement with Tencent. She is currently a Special Associate Researcher with USTC. Her research interests include speech enhancement, robust speech recognition, audio-visual scene classification, sound event localization and detection.

Jun Du (Senior Member, IEEE) received the B.Eng. and Ph.D. degrees from the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), Hefei, China, in 2004 and 2009, respectively. From 2009 to 2010, he was with iFLYTEK Research as a Team Leader, working on speech recognition. From 2010 to 2013, he joined Microsoft Research Asia as an Associate Researcher, working on handwriting recognition and OCR. Since 2013, he has been with the National Engineering Laboratory for Speech and Language Information Processing, USTC. He has authored or coauthored more than 150 papers. His main research interests include speech signal processing and pattern recognition applications. He is an Associate Editor for IEEE/ACM TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING and a member of the IEEE Speech and Language Processing Technical Committee. He was the recipient of the 2018 IEEE Signal Processing Society Best Paper Award. His team won several champions of the CHiME-4/CHiME-5/CHiME-6 Challenge, SELD Task of the 2020 DCASE Challenge, and the DIHARD-III Challenge.

Sabato Marco Siniscalchi (Senior Member, IEEE) received the Laurea and Doctorate degrees in computer engineering from the University of Palermo, Palermo, Italy, in 2001 and 2006, respectively. In 2001, he was with STMicroelectronics where he designed optimization algorithms for processing digital image sequences on very long instruction word (VLIW) architectures. In 2002, he was an Adjunct Professor with the University of Palermo and taught several undergraduate courses for computer and telecommunication engineering. In 2006, he was a Postdoctoral Fellow with the Center for Signal and Image Processing (CSIP), Georgia Institute of Technology, Atlanta, GA, USA, under the guidance of Prof. C.-H. Lee. From 2007 to 2009, he was with the Norwegian University of Science and Technology, Trondheim, Norway, as a Research Scientist, Department of Electronics and Telecommunications under the guidance of Prof. T. Svendsen. In 2010, he was a Researcher Scientist with the Department of Computer Engineering, University of Palermo. He is currently an Assistant Professor with the University of Enna “Kore,” Enna, Italy. He has coauthored more than 20 papers in this field. His main research focuses on speech processing, in particular automatic speech and speaker recognition, and language identification.

Shi-Fu Xiong received the Bachelor of Science degree from the School of Electronic Information and Communication, Huazhong University of Science and Technology, Wuhan, China, in 2011, and the master of Science degree from the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, China, in 2014. Since 2014, he has been with iFLYTEK Research on spoken dialog systems. His research focuses on speech recognition and natural language processing.

Gen-Shun Wan received the B.S. and M.S. degrees from Jiangsu University, Zhenjiang, China, in 2012 and 2015, respectively. He is currently working toward the Ph.D. degree with the University of Science and Technology of China, Hefei, China. Since 2015, he has been performing iFlytek Research on speech recognition and spoken dialog systems. His research interests include speech recognition, dialog systems, and machine learning. He won first place in many speech challenges, including OpenASR Challenge 2021 and the CHiME Challenge 7.