Computer Speech and Language 99 (2026) 101924

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

Audiovisual speech enhancement and voice activity detection using generative and regressive visual features

Cheng Yu [a,*], Vahid Ahmadi Kalkhorani [a], Buye Xu [b], DeLiang Wang [c]

[a] Department of Computer Science and Engineering, Ohio State University, Columbus, OH, 43210, USA
[b] Reality Labs, Meta, Redmond, WA, 20004, USA
[c] School of Data Science, The Chinese University of Hong Kong, Shenzhen, Guangdong, 518172, China

ARTICLE INFO

Keywords:
Speech enhancement
Audiovisual speech enhancement
Voice activity detection
Multi-task learning

ABSTRACT

We present an audiovisual speech enhancement (AVSE) system to address two related tasks: speech enhancement (SE) and voice activity detection (VAD). The system is based on a complex spectral mapping model and performs two-stage audiovisual fusion. The first stage is a signal-level fusion module, where a generative lip-to-speech conversion method produces time-frequency (T-F) features from lip movements. This allows the system to leverage noise-free T-F representations, which are crucial for improving speech intelligibility, particularly in challenging acoustic environments. The second stage is an embedding-level fusion module, where high-dimensional embedding features from a jointly trained visual encoder are integrated. Additionally, we propose a multitask learning framework that optimizes both SE and VAD tasks. The inclusion of a VAD decoder enables the system to distinguish speech from non-speech segments. We evaluate the system on multiple benchmark datasets, including COG-MHEAR, LRS3-AudioSet, and LRS3-CHiME3, and achieve state-of-the-art SE and speech recognition results, and significant robustness in VAD compared to the audio-only baseline. These results highlight the effectiveness of our system in realistic environments.

1. Introduction

Speech enhancement (SE) aims to enhance the quality and intelligibility of speech corrupted by background noise and room reverberation (Loizou, 2013). During the past dozen years, SE algorithms have seen large advances, driven by the introduction of deep learning methods (Wang and Chen, 2018). Despite these achievements, SE algorithms still suffer from notable degradation when applied to adverse acoustic scenarios, and improving SE algorithms to perform well in highly noisy conditions continues to be a difficult task (Hao et al., 2020; Wang and Wang, 2023). In addition to exploring solutions within the audio modality itself, researchers have investigated combining auditory and visual modalities to address these limitations (Michelsanti et al., 2021). Since auditory and visual cues are complementary in human speech perception (Schwartz et al., 2004), leveraging these modalities offers potential advantages. In particular, the visual modality is immune to acoustic interference, benefiting a range of speech processing tasks in background interferences, such as voice activity detection (VAD) (Shahid et al., 2021), automatic speech recognition (ASR) (Ma et al., 2022), and other applications (Prajwal et al., 2020b). This unique property has inspired the development of audiovisual speech enhancement (AVSE) that exploits both modalities to enhance the robustness of SE algorithms (Hou et al., 2018; Afouras et al., 2018b).

* Corresponding author.
E-mail address: yu.3500@osu.edu (C. Yu).

https://doi.org/10.1016/j.csl.2025.101924
Received 6 May 2025; Received in revised form 31 October 2025; Accepted 9 December 2025
Available online 14 December 2025

There are various approaches to AVSE (Michelsanti et al., 2021). These algorithms utilize informative visual features, such as mouth images (Hou et al., 2018), facial landmarks (Morrone et al., 2019), and visual speech or speaker recognition embeddings (Afouras et al., 2018b; Sun et al., 2020). There are generally two audiovisual fusion strategies, masking (Afouras et al., 2018b) and concatenation (Hou et al., 2018). These early fusion studies have reported notable improvements over audio-only systems. However, they are generally based on a regressive model structure and have limited performance in more challenging noise conditions (Mira et al., 2023). To address these challenges, some approaches propose resynthesis-based algorithms that first generate acoustic units, such as Mel spectrograms (Mira et al., 2023) or speech codecs (Yang et al., 2022), and then use neural decoders to reconstruct the enhanced features into audio waveforms. These methods have proven to be effective in generating quality speech signals in extremely adverse conditions, such as with multiple interferences and very low signal-to-noise ratio (SNR) levels (Mira et al., 2023; Yang et al., 2022). However, the signals generated by these methods require complex two-stage encoder–decoder training. Moreover, these algorithms heavily rely on audio decoders, which means that their performance is constrained by the efficiency and reliability of neural vocoders, which can be computationally expensive and are not very reliable when evaluated with perceptual metrics (Liao et al., 2024). Others employ generative algorithms, such as variational autoencoders (Sadeghi et al., 2020; Sadeghi and Alameda-Pineda, 2021), diffusion-based models (Richter et al., 2023a), and adversarial generative networks (Xu et al., 2022). Despite good generalizability on untrained data, unsupervised algorithms are typically outperformed by supervised systems under similar training and test conditions (Richter et al., 2023b).

Recently, advanced regressive AVSE (Pan et al., 2023) and audiovisual speaker separation (AVSS) (Kalkhorani et al., 2024, 2025) approaches have shown superior performance compared to strong baselines. These methods often leverage visual features extracted by a pre-trained ResNet-based visual encoder (He et al., 2016), originally designed for audiovisual speech recognition (AVSR) tasks (Afouras et al., 2018a). By incorporating visual embeddings, these algorithms feature powerful regressive architectures (Wang et al., 2023; Kalkhorani and Wang, 2024), which offer several advantages over more generative methods; for example, regressive models are generally easier to train, and their optimization can be more directly guided by evaluation metrics. Despite their impressive performance, it is unclear whether these systems offer any advantages over their audio-only counterparts in the context of SE tasks (Pan et al., 2023). So an open research question is how an AVSE system can benefit from visual features when it is built on a state-of-the-art audio-based model.

In addition to AVSE, recent advances in cross-modal generative algorithms have garnered significant attention. In particular, algorithms that convert video to audio (Prajwal et al., 2020a; Hsu et al., 2023; Yemini et al., 2024) have potential benefits for AVSE systems. These methods are inherently insensitive to acoustic interference and can be trained on large-scale datasets independently of specific downstream tasks. Although such generative models have shown promising results in speech synthesis, the synthesized audio still does not fully match the natural characteristics of human speech at the signal level (Yemini et al., 2024). This limitation raises important questions about the potential of these methods for real-world applications where natural-sounding, individualized speech is needed. Nonetheless, progress in cross-modal synthesis underscores the growing intersection between vision and audition, potentially opening new avenues for improving AVSE and related tasks.

In this work,[1] we conduct a comprehensive study on the impact of visual cues on the overall performance of audiovisual speech enhancement, specifically in the context of state-of-the-art audio-only SE architecture. To take advantage of state-of-the-art regressive SE architecture and noise insensitivity of the visual modality, we propose a regressive AVSE model on the basis of an audio-only system, and the model is designed to benefit from two kinds of visual feature. Specifically, the proposed AVSE model incorporates both generative and regressive visual features through a novel two-stage fusion strategy. In the first, signal-level fusion module, we integrate Mel spectrograms from the diffusion-based generative LipVoicer model (Yemini et al., 2024) and complex spectrograms from acoustic input signals before feeding them into a convolutional encoder. In the second, embedding-level fusion module, we use a FaceEncoder to extract visual embeddings, which are combined with the output of signal-level fusion.

Our proposed model is designed to perform simultaneously the SE and noise-robust VAD tasks. By jointly optimizing these tasks, the model benefits from shared representations, improving the accuracy of speech detection while enhancing the quality of the speech signal. The VAD can function either as an independent task or as a postprocessor to the enhanced speech. VAD-based postprocessing significantly enhances the quality of the processed speech by providing better noise suppression in silent frames. Multitask learning has been explored in audio-only settings, which leverages SE to support VAD in noisy conditions (Zhuang et al., 2016; Nonaka et al., 2021; Tan and Zhang, 2021). To our knowledge, the proposed model is the first deep learning-based audiovisual model capable of simultaneously performing SE and VAD tasks. Experimental results demonstrate that the proposed model achieves state-of-the-art performance on the COG-MHEAR AVSE challenge, the LRS3-AudioSet, and the LRS3-CHiME3 dataset, along with superior VAD results.

2. Model description

In this section, we first formulate the research problem and introduce the proposed model architecture. Then we describe the two visual encoders in the model, the two-stage audiovisual fusion strategy, and multitask learning. The proposed AVSE system adopts both generative- and regressive-based visual features. Given a recorded visual stream $\mathbf{V}$ and the noisy speech signal of a single speaker $y(t)$ with the voice activity of $d(t)$, we formulate $y(t)$ in terms of the target speech signal $x(t)$ and additive noise signal $n(t)$ as

$$y(t) = x(t) + n(t), \tag{1}$$

[1] All research and experiments were conducted at the Ohio State University.

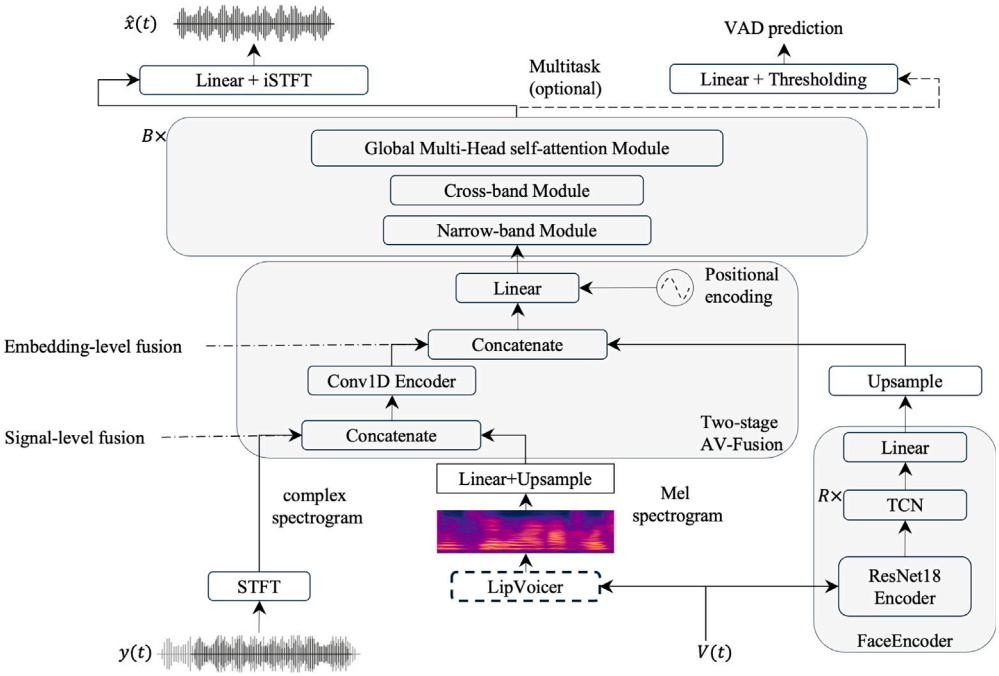

Fig. 1. Proposed system architecture. Visual cues **V** are fed to LipVoicer and the ResNet-based visual encoder to generate two levels of visual features for two-stage audiovisual fusion. The dash-boxed LipVoicer indicates that its parameters are fixed during training and test.

where t denotes the time index of the waveform signal. The above time-domain formulation can be expressed in the short-time Fourier transform (STFT) domain as follows:

$$\mathbf{Y}(m, f) = \mathbf{X}(m, f) + \mathbf{N}(m, f), \tag{2}$$

$$\mathbf{Y}(m, f) = Y_r(m, f) + iY_i(m, f). \tag{3}$$

where m and f index time frame and frequency bin respectively. The proposed AVSE system takes time-aligned **Y** and **V** as inputs and generates the enhanced speech, $\hat{\mathbf{X}}$, and an optional voice activity estimate, $\hat{\mathbf{d}}$, as

$$\hat{\mathbf{X}},\ \hat{\mathbf{d}} = \text{AVSE}(\mathbf{Y},\ \mathbf{V}). \tag{4}$$

The proposed system is designed to strike a balance between generative and regressive visual encoders. To do so, we propose two-stage audiovisual fusion that utilizes generative and regressive features at two separate stages. Fig. 1 illustrates the proposed model architecture, which consists of a two-stage audiovisual fusion module, a scalable number (B) of TF-CrossNet blocks (Kalkhorani and Wang, 2024), and a pair of linear layers that serve as the SE and VAD decoders. The two-stage audiovisual fusion module processes three kinds of input: the audio complex spectrogram feature, generative visual features from the LipVoicer model, and regressive visual embeddings from the ResNet-based visual encoder.

2.1. Visual features

2.1.1. Generative features

For generative visual features, we adopt the Mel spectrogram generated by the LipVoicer encoder (Yemini et al., 2024). LipVoicer is a diffusion-based generative model designed for modality transformation, specifically converting video into the Mel spectrogram, which is then synthesized to speech using a vocoder. It is capable of reconstructing speech from silent lip movements. To ensure proper alignment between auditory and visual content, LipVoicer is designed to synchronize the lip-reading results and the generated speech output, aligning them with the predictions of a pre-trained lip-reading model and an ASR model during inference. This process is guided by matching the recognition results of the predicted Mel spectrogram at each diffusion step with the corresponding lip reading output. Furthermore, to preserve the speaker's identity in the generated features, the diffusion process is conditioned on a frontal image of the target speaker's face.

2.1.2. Regressive features

For regressive visual features, we adopt a visual encoder based on the ResNet18 (He et al., 2016) architecture, which has been successfully applied in various visual and audiovisual tasks (Afouras et al., 2018a; Tesema et al., 2023). Unlike previous works that

directly use visual embeddings pre-extracted from the AVSR visual encoder (Afouras et al., 2018a), we have the flexibility to either initialize our visual encoder with pre-trained weights or train it from scratch. We initialize the encoder with pre-trained weights from (Afouras et al., 2018a) and fine-tune it on our task to further improve performance. Jointly training the visual encoder and our model allows it to generate more task-specific regressive features, leading to improved performance.

2.2. Two-stage audiovisual fusion

As illustrated in Fig. 1, the proposed fusion module encodes the visual cues $\mathbf{V}$ from two encoders, and performs audiovisual fusion at two different levels. For the signal-level fusion, we apply two steps to the Mel spectrogram $\mathbf{M}$ generated by LipVoicer. First, the Mel spectrogram is temporally upsampled using linear interpolation to match the frame rate of the complex audio spectrogram. Second, it is processed through a linear layer to expand its frequency dimension, aligning it with that of the audio spectrogram. The steps are formulated as follows:

$$\mathbf{M} = \text{LipVoicer}(\mathbf{V}), \quad \mathbf{M} \in \mathbb{R}^{T_L \times 1 \times 80} \tag{5}$$

$$\mathbf{M}' = \text{Linear}(\text{Upsample}(\mathbf{M})), \quad \mathbf{M}' \in \mathbb{R}^{T \times 1 \times F} \tag{6}$$

where T_L, T, and F represent the temporal dimension of the Mel spectrogram and the temporal and frequency dimensions of the complex spectrogram, respectively. The embedding $\mathbf{M}'$ is then concatenated with the complex spectrogram along the channel dimension, resulting in an input with a channel size of 3, i.e., two channels of complex audio spectrogram and one channel of expanded Mel spectrogram $\mathbf{M}'$. At this stage, both the complex spectrogram Y_r, Y_i, and the processed Mel spectrogram $\mathbf{M}'$ are represented in the time–frequency domain. The signal-level fusion module aims to enhance the robustness of the time–frequency features by leveraging both modalities in an early stage. The combined signal-level features $\mathbf{E1}$ are then passed through the one-dimensional (1-D) convolutional encoder, referred to as the Conv-1D Encoder, which is formulated below:

$$\mathbf{Y}' = \text{Concatenate}(Y_r, Y_i, \mathbf{M}'), \quad \mathbf{Y}' \in \mathbb{R}^{T \times 3 \times F} \tag{7}$$

$$\mathbf{E1} = \text{Conv1D Encoder}(\mathbf{Y}'), \quad \mathbf{E1} \in \mathbb{R}^{T \times C \times F}. \tag{8}$$

where C represents the channel dimension of $\mathbf{E1}$. For the embedding-level fusion, the visual embeddings are generated by the FaceEncoder, which consists of a ResNet18 encoder, five temporal convolutional network (TCN) layers (Lea et al., 2017), and a final linear layer. The ResNet18 encoder is designed to extract facial features from the input video frames. Specifically, it captures facial features in each frame. The TCN captures the temporal dynamics of the visual embeddings extracted by ResNet18, enabling the model to learn from sequential facial movements over time. The linear layer projects these temporal features into the embedding space, facilitating effective fusion with the audio features in the subsequent stages of the system. To ensure proper alignment, the visual embeddings are up-sampled using linear interpolation to match the temporal dimension of the embeddings $\mathbf{E1}$ from the Conv1D Encoder. This process is formulated as follows:

$$\mathbf{V}' = \text{FaceEncoder}(\mathbf{V}), \quad \mathbf{V} \in \mathbb{R}^{T_v \times C \times F} \tag{9}$$

$$\mathbf{E2} = \text{Upsample}(\mathbf{V}'), \quad \mathbf{E2} \in \mathbb{R}^{T \times C \times F}, \tag{10}$$

where T_v represents the temporal dimension of the video. Next, we concatenate the embeddings $\mathbf{E1}$ and $\mathbf{E2}$ along the channel axis to form the joint embeddings $\mathbf{E3}$, which is then passed through a linear layer to restore the original channel shape. This embedding-level fusion combines the two embeddings within a shared high-dimensional space, enabling the model to harness the benefits of two distinct visual feature streams. More specifically, the LipVoicer stream strengthens low-level features by using signal-level fusion module in the first stage. The FaceEncoder stream provides more abstract visual embeddings through the embedding-level fusion module in the second stage. The two-stage fusion allows the model to integrate complementary information from two streams. The second-stage fusion is formulated as follows:

$$\mathbf{E3} = \text{Concatenate}(\mathbf{E1}, \mathbf{E2}), \quad \mathbf{E3} \in \mathbb{R}^{T \times 2C \times F}, \tag{11}$$

$$\mathbf{E} = \text{Linear}(\mathbf{E3}), \quad \mathbf{E} \in \mathbb{R}^{T \times C \times F}. \tag{12}$$

The high-dimensional feature $\mathbf{E}$ is then fed to the TF-CrossNet blocks after adding random-chunk positional encoding (Kalkhorani and Wang, 2024).

Before reaching the above design of two-stage audiovisual fusion, we explored several other techniques, including just the early fusion of LipVoicer features, which improved speech intelligibility metrics at low SNRs but reduced perceptual quality metrics at higher SNRs. To better align with LipVoicer outputs, we also adapted a model to process noisy Mel-spectrogram audio inputs and used a HiFi-GAN vocoder to reconstruct the waveforms from the enhanced audio features, following a related fusion strategy that combines noisy and generative Mel representations from synchronized lip videos (Gabbay et al., 2018). Although this Mel-based design showed a similar trend of results, its performance is limited by the vocoder's synthesis quality. Building on these observations, the proposed signal-level fusion integrates LipVoicer Mel features and complex spectrogram inputs, achieving a better balance between intelligibility and quality metrics. The embedding-level fusion further fine-tunes the results using regressive visual embeddings.

2.3. TF-CrossNet block

A TF-CrossNet block (Kalkhorani and Wang, 2024) consists of three modules: Narrow-band module, cross-band module, and global multi-head self-attention (GMHSA) module.

2.3.1. Narrow-band module

The narrow-band module consists of layer normalization, a linear layer with SiLU (sigmoid linear unit) activation, a temporal-convolutional layer (T-Conv), and a final linear layer. The first linear layer increases the feature dimension, while the last one restores it to its original size. The T-Conv layer includes grouped 1D convolutions with SiLU activations, and a grouped normalization layer after the second convolution. This module is a simplified version of the Conformer convolutional block (Gulati et al., 2020), with the MHA module removed, as narrow-band correlations are already captured by the GMHSA module in TF-CrossNet.

2.3.2. Cross-band module

This frequency-convolutional module is designed to capture the correlations between neighboring frequencies. It consists of a grouped convolution, followed by layer normalization and PReLU (parametric rectified linear unit) activation (He et al., 2015). In the full-band linear submodule, we first apply a linear layer with SiLU activation (Hendrycks and Gimpel, 2016). Then, a series of channel-independent linear layers operates along the frequency axis to extract full-band features. To be more specific, each linear layer is dedicated to its assigned feature channel. Finally, the output of the module is generated by restoring the channels back to their original shape through a linear layer with the SiLU activation, which is then added to the initial input of the module.

2.3.3. Global multi-head self-attention

The GMHSA module applies a self-attention layer to the embeddings to capture global correlations. The results of all attention heads are concatenated and passed through a point-wise convolution, followed by a PReLU activation function and layer normalization. A residual connection is incorporated between the input and output to refine the final result. Additionally, the GMHSA module merges all frequency features into the channel dimension before applying multi-head self-attention. This approach allows each frame of embeddings to attend to other frames across all feature channels, capturing long-range correlations in both frequency and hidden feature channels while maintaining computational efficiency.

2.4. Multitask decoders

As illustrated in Fig. 1, the proposed AVSE system can be trained in two configurations: SE alone, or multitask learning that simultaneously handles SE and VAD. In both configurations, the SE decoder generates the enhanced complex audio spectrogram of target speech, which is then converted back into the time domain using the inverse short-time Fourier transform (iSTFT). The multitask version only requires an additional linear decoder, making it computationally efficient. In this case, the VAD task shares a common encoder with the SE task, enabling the system to leverage shared feature representations for both tasks. The features extracted from the final TF-CrossNet block are passed to both the SE and VAD decoders. Meanwhile, the VAD decoder outputs speech activity predictions $\hat{\mathbf{d}}$, which contain a single value for each time frame, representing the likelihood of speech activity. To obtain the final VAD output, thresholding is applied based on the speech presence ratio. This process enables the model to effectively classify each frame as either speech or non-speech, allowing for improved noise suppression and speech activity detection in noisy environments.

2.5. Loss functions

We train all systems using a hybrid loss (Pan et al., 2022a), which consists of two components: the scale-invariant signal-to-distortion ratio (SI-SDR) loss and a multi-resolution STFT loss. The SI-SDR loss, or $L_{\text{SI-SDR}}$, is applied in its standard form, where the target signal is scaled to match the estimated signal. For the multi-resolution STFT loss, $\mathcal{L}_{\text{multiSTFT}}$, we adopt three window lengths of $256, 512, 1024$, with the corresponding window shifts of $25, 60, 120$. To train the multitask model, we incorporate an additional binary cross entropy (BCE) loss into the hybrid loss. The BCE loss is defined as:

$$\mathcal{L}_{\text{BCE}} = \frac{1}{T}\sum_{m=1}^{T}[d(m)\cdot\log(\hat{d}(m)) + (1-d(m))\cdot\log(1-\hat{d}(m))] \tag{13}$$

where $d(m)$ and $\hat{d}(m)$ denote target label and predicted VAD label respectively, with m indexing time frame. The overall loss functions can be defined as:

$$\mathcal{L}_{\text{hybrid}} = \mathcal{L}_{\text{SI-SDR}} + \mathcal{L}_{\text{multiSTFT}} \tag{14}$$

$$\mathcal{L}_{\text{multitask}} = \mathcal{L}_{\text{SI-SDR}} + \mathcal{L}_{\text{multiSTFT}} + \mathcal{L}_{\text{BCE}} \tag{15}$$

where $\mathcal{L}_{\text{hybrid}}$ is adopted for AVSE, and $\mathcal{L}_{\text{multitask}}$ is adopted for multitask learning.

3. Experimental setup

3.1. Datasets

We use three datasets to evaluate the proposed AVSE system. These datasets are the COG-MHEAR challenge (Blanco et al., 2023), the LRS3-AudioSet, and the LRS3-CHiME3 (Richter et al., 2023a).

3.1.1. COG-MHEAR

The COG-MHEAR dataset is a benchmark provided by the COG-MHEAR AVSE challenge (Blanco et al., 2023). The audiovisual signals in this dataset are sourced from the LRS3 dataset (Afouras et al., 2018c), originally designed for visual speech recognition, and are temporally aligned to serve as target signals for the AVSE task. The accompanying noise signals are collected from three sources: domestic noises from the Clarity Challenge (Graetzer et al., 2021), the Demand noise set (Thiemann et al., 2013), and various noise clips from the Freesound database (Fonseca et al., 2017), as part of the second DNS challenge (Reddy et al., 2021). The dataset includes two main tasks: AVSE and audiovisual target speaker extraction (AVTSE). As this work focuses on AVSE, we evaluate using only the AVSE portion of the dataset. Since the ground truth signals for the test dataset are not available, we reorganize the training set into training and validation subsets while treating the development set as the test set, similar to Pan et al. (2023), Kalkhorani et al. (2024). To create the new validation dataset, we select utterances from 24 speakers in the COG-MHEAR training set, prioritizing those with the largest number of utterances, resulting in 3017 utterances. The remaining data is used for training. Consequently, the total number of utterances in the reorganized training, validation, and test sets are 30,297, 3017, and 1661, respectively. Both training and validation utterances are set to a consistent length of 3.0 s, while the test utterances vary in duration, ranging from approximately two seconds to over thirty seconds. The dataset uses speech-weighted signal-to-noise ratios (SNRs) randomly drawn from the range of $[-10,\ 10]$ dB, simulating various noisy conditions commonly encountered in real-world scenarios.

3.1.2. LRS3-AudioSet

Similar to other AVSE studies (Hsu et al., 2023; Jung et al., 2024), we prepare the LRS3-AudioSet dataset as follows. The target audio signals are sourced from the trainval set of the original LRS3 dataset, which consist of around 100k, 10k, and 2.1k utterances for training, validation, and test, respectively. The noise signals are sourced from the AudioSet non-speech subset, which consists of 2.7k, 305, and 86 signals for training, validation, and test, respectively. We randomly mix the utterances and non-speech signals using the standard SNR selected from $[-15,\ 15]$ dB. For VAD experiments, we create more challenging mixtures in the SNR range of $[-35,\ 15]$ dB to evaluate the effectiveness of the proposed model.

3.1.3. LRS3-CHiME3

For the LRS3-CHiME3 dataset, we assess the generalizability of our model using its test set. Following the procedure outlined in Richter et al. (2023b), we select 244 utterances from the LRS3 test set, ensuring that each utterance is at least 3.2 s in duration. These utterances are randomly mixed with noise signals from the CHiME3 dataset (Barker et al., 2015) in the standard SNR range of $[-6,\ 12]$ dB.

3.2. Hyperparameters

To ensure a consistent experimental setup, all audio signals are resampled to 16 kHz, while all videos are resampled to 25 frames per second (FPS). For audio features, we apply STFT with a window size of 256 samples and a shift size of 128, resulting in 129 frequency bins per frame. For visual features, the LipVoicer generates a Mel spectrogram with 80 frequency bins and a 160 sample shift size. The ResNet18 encoder outputs a 512-dimensional feature vector with a shift size of 640 along the audio time axis. To ensure alignment, we upsample visual features to match audio feature shapes, using interpolation for temporal alignment and linear layers to adjust the time–frequency resolution. To prepare the VAD configuration, we apply the signal-processing-based rVAD (Tan et al., 2020) algorithm to clean speech signals to generate the target VAD labels. A threshold of 0.85, based on the speech presence rate on the validation set, is applied to the proposed system and all the baselines. In the model architecture shown in Fig. 1, we set the number of TF-CrossNet blocks $B = 12$ and $B = 6$ for full-size and half-size models, respectively. The number of TCN layers R is set to 5 for all configurations. The kernel sizes of the audio encoder layer, T-Conv layers, and frequency group convolutional layers are set to 5, 5, and 3, respectively. The number of groups for T-Conv layers, frequency group convolutional layers, and group normalization is all set to 8. The hidden channel sizes are set to $H = 192$, $H' = 16$, and $H'' = 384$ according to the original paper (Kalkhorani and Wang, 2024). We employ 4 self-attention heads in the GMHSA module with an embedding dimension of 5.

To ensure efficient training, we employ mixed-precision (bf16) training. For consistent data batch generation, we truncate both training and validation utterances to a uniform length of 3.0 s per chunk. We also utilize the maximum batch size that fits within GPU memory constraints, approximately 46 GB on an Nvidia A40. Training is stopped when no improvement in validation loss is observed for 10 consecutive epochs. We use the Adam optimizer with an initial learning rate of 10^{-3} and apply the PyTorch ReduceLROnPlateau scheduler to adjust the learning rate. The scheduler is configured with a patience of 3 epochs and a reduction factor of 0.85.

3.3. Baselines

We choose several audio-only and audiovisual models as baselines for quantitative comparisons. For audio-only models, we train publicly available TF-CrossNet (Kalkhorani and Wang, 2024) and HD-DEMUCS (Kim et al., 2023) to assess the impact of the visual modality. TF-CrossNet is based on complex spectral mapping, while HD-DEMUCS is based on waveform mapping. Note that HD-DEMUCS is the current state-of-the-art variant of the DEMUCS model (Defossez et al., 2020) for SE. To examine the role of audiovisual signal-level fusion, we trained an audiovisual variant of HD-DEMUCS, the AV-HD-DEMUCS model, using this fusion approach. We want to emphasize that training an audiovisual variant of audio-only SE models as baseline models is widely adopted in AVSE studies (Mira et al., 2023; Zhu et al., 2023; Chen et al., 2024). Furthermore, we select the high-performing AVDPRNN and AVGridNet models (Pan et al., 2023), following the procedures outlined in the original papers and using the publicly available code.[2] These models employ pre-extracted DeepAVSR (Afouras et al., 2018a) features via embedding-level fusion. In addition, we train a multitask version of TF-CrossNet to serve as an audio-only multitask baseline. To evaluate VAD performance, we consider two recent, strong methods, which produce comparable results under similar training and test conditions (Zhao and Champagne, 2022; Tang et al., 2024). As a baseline, we select the publicly available model from (Zhao and Champagne, 2022), referred to as Tr-VAD. To assess the contribution of visual features, we also create a video-only system that utilizes the same architecture as the proposed system but processes visual inputs only.

3.4. Evaluation metrics

We use several widely used objective metrics. For the AVSE task, we measure performance using the wide-band perceptual evaluation of speech quality (PESQ) (Thiemann et al., 2013), which ranges from −0.5 to 4.5, with higher scores indicating better speech quality. We also use standard or extended short-time objective intelligibility (STOI) (Taal et al., 2010; Jensen and Taal, 2016) depending on the benchmark, which ranges typically from 0.0 to 1.0, roughly corresponding to percent-correct speech intelligibility score. Furthermore, we employ SI-SDR measured in dB. For the LRS3-AudioSet benchmark, we additionally measure performance using three perceptual metrics based on Mean Opinion Score (MOS) prediction (Hu and Loizou, 2008): signal distortion (CSIG), background intrusiveness (CBAK), and overall quality (COVL). These metrics range from 1 to 5, with higher scores indicating better perceptual quality.

For the VAD task, we use the F1 score, which balances precision and recall. It can be formulated as:

$$\mathrm{F1} = \frac{2}{1/Precision + 1/Recall} \tag{16}$$

where $Precision = \mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$, $Recall = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$, and TP, FP, and FN denote true positives, false positives, and false negatives, respectively. The F1 score is the harmonic mean of Precision and Recall. Specifically, it considers both FP (incorrectly detecting speech) and FN (failing to detect speech), to ensure the accuracy and reliability of a VAD system. The F1 score ranges from 0.0 to 1.0, with 1.0 indicating perfect classification. To provide a deeper understanding, we also evaluate VAD using the Hit-FA rate (Kim et al., 2009). The Hit-FA rate combines the Hit rate (true positive rate) and the FA (false-alarm or false positive) rate. It penalizes algorithms with high FA rates, which are known to have a major negative impact on speech intelligibility.

Furthermore, we measure word error rate (WER), a standard ASR evaluation metric, of enhanced speech. The WER is calculated using state-of-the-art ASR toolkits. Specifically, we adopt the QuartzNet15 × 5Base-En checkpoint from NeMo (NVIDIA, 2021), and the medium.en checkpoint from Whisper (Radford et al., 2022). For the LRS3-AudioSet dataset, we follow the standard WER computation, which accounts for the total number of substitutions, insertions, and deletions under each test condition. As the original transcriptions are not provided with the dataset, we generate the ground truth transcriptions using the clean target audio. Both the ground truth and predicted transcriptions are obtained using the same Whisper model. Notably, the adopted Whisper model achieves a WER of 2.3% on the LRS3 test set, demonstrating its transcription accuracy on clean speech (Rouditchenko et al., 2024). For the LRS3-CHiME3 dataset, we follow the baseline approach by generating transcriptions using the NeMo model and computing the average WER across all audio samples (Richter et al., 2023a). The WER is calculated using the ground truth transcriptions provided with the dataset.

4. Experimental results

4.1. COG-MHEAR challenge dataset

Table 1 presents the SE evaluation and comparison results on the COG-MHEAR challenge dataset. The proposed system outperforms all other approaches across all metrics. Specifically, it achieves an improvement of 0.11 and 0.2 in PESQ and SI-SDR over the state-of-the-art audio-only model, TF-CrossNet, and a 0.05 PESQ gain over AV-CrossNet (Kalkhorani et al., 2025). The improvement over AV-CrossNet can be attributed to two factors. First, while AV-CrossNet relies on pre-extracted visual embeddings from a pre-trained visual encoder, our system trains the ResNet18 visual encoder as part of the integrated system (see Fig. 1). Second, we include generative features through signal-level fusion. It is worth noting that, at its publication in 2023, SAV-GridNet achieved the state-of-the-art performance, with a PESQ of 2.68 and an SI-SDR of 14.2 dB, substantially outperforming other AV baselines on

[2] https://github.com/zexupan/avse_hybrid_loss

Table 1
Speech enhancement results on the COG-MHEAR AVSE Challenge dataset (AVSE part).

Method	AV	Params (M)	PESQ↑	STOI↑	SI-SDR↑
Unprocessed	–	–	1.15	0.68	−4.4
AV-DPRNN (Pan et al., 2023)	✓	4.1	2.02	0.86	11.4
SAV-GridNet (Pan et al., 2023)	✓	21.5	2.68	0.91	14.2
TF-CrossNet (Kalkhorani and Wang, 2024)	✗	7.3	2.69	0.91	14.3
AV-CrossNet (Kalkhorani et al., 2025)	✓	11.1	2.75	**0.92**	14.3
Proposed	✓	9.6	**2.80**	**0.92**	**14.5**

Table 2
Speech enhancement results on LRS3-AudioSet.

Method	SE metrics↑					
	PESQ	STOI	SI-SDR	CSIG	CBAK	COVL
Unprocessed	1.32	0.72	0.08	2.29	1.90	1.74
Audio-only methods						
HD-DEMUCS (Kim et al., 2023)	2.25	0.83	10.7	3.53	2.92	2.88
TF-CrossNet (Kalkhorani and Wang, 2024)	2.77	**0.90**	**13.9**	4.04	3.40	3.44
TF-CrossNet[a] (Kalkhorani and Wang, 2024)	**2.79**	**0.90**	13.6	**4.09**	**3.43**	**3.48**
Audiovisual methods						
AV-HD-DEMUCS (Kim et al., 2023)	2.35	0.87	11.8	3.70	3.07	3.03
AV-DPRNN (Pan et al., 2023)	2.33	0.88	12.4	3.70	3.09	3.01
AV-GridNet$_n$ (Pan et al., 2023)	2.81	0.92	13.2	4.16	3.49	3.53
AV-CrossNet (Kalkhorani et al., 2025)	2.81	0.92	13.5	4.15	3.44	3.51
Proposed	2.87	**0.93**	**13.9**	**4.21**	**3.53**	**3.59**
Proposed[a]	**2.88**	0.92	13.8	4.18	3.49	3.56

[a] Indicates multitask configuration.

the COG-MHEAR challenge. Compared to SAV-GridNet, we achieve a 0.12 improvement in PESQ and a 0.3 dB increase in SI-SDR. These results clearly demonstrate the superior performance, advancing the state-of-the-art on the AVSE task. The column of Params (M) reports the number of parameters in millions in the SE module. All the AV methods listed employ a ResNet18-based encoder, which has 11.2M parameters. The proposed system uses a smaller SE module compared to AV-CrossNet, owing to the exclusion of the narrowband MHA module. In contrast, SAV-GridNet incorporates two expert models (9M parameters each) along with a scenario classifier, resulting in substantially more parameters. Additionally, the proposed method includes a LipVoicer model with 57.3M parameters, contributing to a higher overall parameter count.

4.2. LRS3-AudioSet dataset

Table 2 and Fig. 2 present the evaluation results on the LRS3-AudioSet dataset, comparing our proposed systems with the audio-only baselines of HD-DEMUCS, TF-CrossNet and TF-CrossNet*, where * denotes multitask configuration, and other AVSE systems. We evaluate these models using the three SE standard metrics of PESQ, STOI, and SI-SDR, the perceptual metrics of CSIG, CBAK, and COVL, and the WER metric, at seven SNR levels (−15 to 15 dB, in 5 dB increments). In terms of SE metrics, the proposed model in multitask configuration outperforms the AV-CrossNet (Kalkhorani et al., 2025) and AV-GridNet$_n$ (Pan et al., 2023) baselines by 0.07 in PESQ, and by 0.6 dB and 0.3 dB in SI-SDR, respectively. The STOI scores are close, as improving STOI on top of high values (above 90%) is not easy due to ceiling effects and is less meaningful as the enhanced speech is almost fully intelligible. Compared to the strong baseline of audio-only TF-CrossNet*, our proposed model improves PESQ by 0.09 and STOI by 0.02, highlighting the utility of audiovisual integration. In terms of the perceptual metrics of CSIG, CBAK, and COVL, the proposed model achieves the highest scores across all three, with CSIG of 4.21, CBAK of 3.53, and COVL of 3.59, indicating better perceptual quality compared to AV-CrossNet and AV-GridNet$_n$.

On the impact of the multitask configuration on TF-CrossNet and the proposed system, the results in Table 2 demonstrate that the proposed model outperforms TF-CrossNet in both configurations across all metrics. The multitask configuration improves TF-CrossNet* on all metrics except for SI-SDR, which slightly decreases from 13.9 to 13.6. A similar pattern is observed for the proposed system, with a marginal gain in PESQ (2.88 vs. 2.87) and a slight reduction in SI-SDR (13.8 vs. 13.9). This may reflect different priorities in the multitask setup. While some studies treat SE as a supporting task for VAD (Tan and Zhang, 2021), our model prioritizes SE with the loss function emphasizing perceptual quality. Consequently, this leads to improvements in perceptual quality metrics, including PESQ, at the expense of signal fidelity, consistent with the trade-offs reported in Tan and Zhang (2021). The limited gain observed for Proposed* likely stems from the visual modality's voice activity cues, reducing reliance on VAD.

The WER results in Fig. 2 are largely consistent with the SE results, demonstrating the effectiveness of our proposed system. As mentioned in Section 3.4, we treat Whisper-transcribed target speech as the ground truth, i.e., the target speech's WER is 0.0%. On average, our system outperforms AV-CrossNet, AV-GridNet$_n$ and audio-only TF-CrossNet* respectively by 7.1%, 10.1% and 35.6% relatively. At the high SNR levels of 10 dB and 15 dB, the top-performing systems are AV-DPRNN and AV-GridNet$_n$, and our proposed

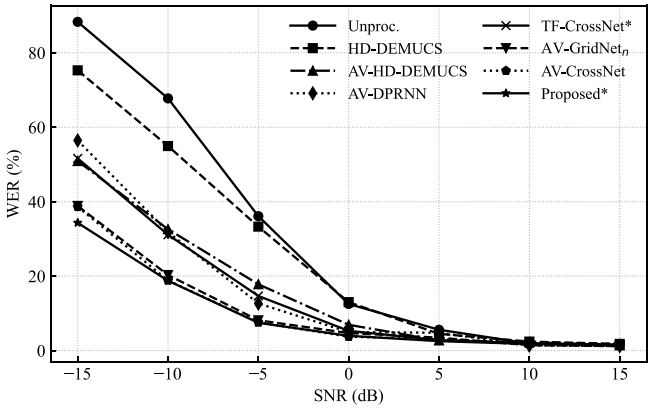

Fig. 2. WER(%) across SNR levels for different AV and audio-only speech enhancement methods. Average WERs are 30.54, 26.47, 16.39, 16.19, 15.57, 11.16, 10.80, and 10.03 for Unprocessed, HD-DEMUCS, AV-HD-DEMUCS, AV-DPRNN, TF-CrossNet*, AV-GridNet$_n$, AV-CrossNet, and Proposed*, respectively.

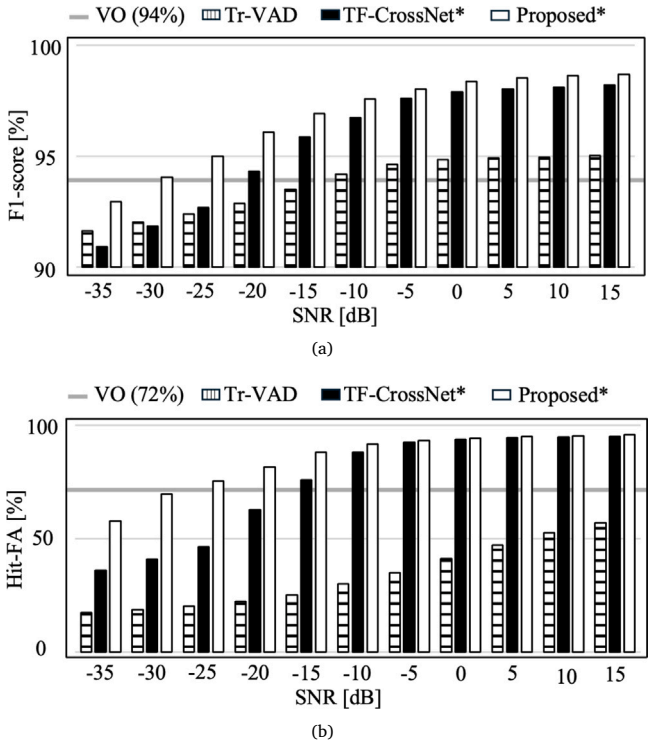

Fig. 3. VAD scores of Proposed* model, TF-CrossNet*, video-only (VO), and Tr-VAD (Zhao and Champagne, 2022) across different SNRs on the test dataset of LRS3-AudioSet. (a) F1 scores. (b) Hit-FA scores.

system shows slightly lower but still competitive WER scores. At lower SNR levels, our system consistently outperforms all the other baselines, with the performance gap widening as SNR decreases. These results highlight the increasing advantages of our audiovisual system as acoustic interference becomes more severe, consistent with human intelligibility data (Sumby and Pollack, 1954).

Fig. 3 presents the VAD results for the multitask configuration of our proposed system, the audio-only TF-CrossNet*, the video-only system (VO), and the baseline model of Tr-VAD (Zhao and Champagne, 2022). We further extend the SNR ranges to extremely low levels to gain additional insights. From Figs. 3(a) and 3(b), it is evident that our proposed AVSE system outperforms both TF-CrossNet* and Tr-VAD at positive SNRs, where the target speech dominates, even though VAD scores are high for all methods. As SNR decreases, the performance gap increases, and becomes particularly noticeable in the Hit-FA results of Fig. 3(b). Tr-VAD shows significantly worse performance in Fig. 3(b), as this metric penalizes the FA rate. As expected, the VO system yields constant

Table 3
Speech enhancement results on the LRS3-CHiME3 dataset.

Method	AV	Params (M)	PESQ↑	ESTOI↑	SI-SDR↑	WER%↓
Unprocessed	–	–	1.20	0.77	2.6	28
SGMSE+ (Richter et al., 2023b)	✗	–	2.08	0.89	10.6	22
HD-DEMUCS (Kim et al., 2023)	✗	23.6	2.11	0.89	11.9	21
VisualVoice (Gao and Grauman, 2021)	✓	54.7	2.11	0.89	11.0	15
AV-Gen (Richter et al., 2023a)	✓	76.0	2.23	0.90	11.4	11
AV-HD-DEMUCS (Kim et al., 2023)	✓	28.7	2.27	0.90	12.5	15
AV-DPRNN (Pan et al., 2022b)	✓	4.1	2.34	0.92	13.6	15
AV-CrossNet (Kalkhorani et al., 2025)	✓	11.1	2.74	**0.94**	13.9	11
AV-GridNet$_n$ (Pan et al., 2023)	✓	9.0	2.79	**0.94**	14.0	**10**
Proposed[a]	✓	9.6	**2.84**	**0.94**	**14.1**	**10**

[a] Indicates the multi-task configuration.

Table 4
Ablation results of different audiovisual fusion strategies on the COG-MHEAR AVSE Challenge dataset (AVSE part).

Method	AV	PESQ↑	STOI↑	SI-SDR↑
Unprocessed	–	1.15	0.68	−4.4
TF-CrossNet (Kalkhorani and Wang, 2024)	✗	2.69	0.91	14.3
AV-CrossNet (Kalkhorani et al., 2025)	✓	2.75	**0.92**	14.3
Signal-level fusion	✓	2.77	**0.92**	14.4
Embedding-level fusion	✓	2.77	**0.92**	**14.5**
Proposed (two-stage fusion)	✓	**2.80**	**0.92**	**14.5**

Table 5
Ablation results of different audiovisual fusion strategies on the LRS3-AudioSet dataset.

Method	AV	PESQ↑	STOI↑	SI-SDR↑
Unprocessed	–	1.32	0.72	0.08
TF-CrossNet (Kalkhorani and Wang, 2024)	✗	2.77	0.90	13.90
AV-CrossNet (Kalkhorani et al., 2025)	✗	2.81	0.92	13.48
Signal-level fusion	✓	2.83	0.92	13.62
Embedding-level fusion	✓	2.84	0.92	13.79
Proposed (two-stage fusion)	✓	**2.87**	**0.93**	**13.93**

performance across different SNRs. The proposed system and TF-CrossNet* outperform the VO baseline at SNRs greater than −15 dB, although the proposed system underperforms VO at the extremely low SNRs of −30 dB and −35 dB; in such conditions, it would not make sense to perform speech processing acoustically. These results show that our proposed system provides robust VAD in very noisy and realistic conditions.

4.3. LRS3-CHiME3 dataset

Table 3 presents the evaluation results on the LRS3-CHiME3 dataset. For this evaluation, we trained our system alongside several baseline systems (AV-CrossNet, AV-GridNet$_n$, AV-DPRNN, AV-HD-DEMUCS, and HD-DEMUCS) using the LRS3-AudioSet dataset. We also compare these systems with those reported in Richter et al. (2023a), which allows us to not only evaluate the performance of our system against existing methods but also test its generalization to CHiME3 noise conditions, which were not seen during training. The proposed system achieves the highest scores on all evaluation metrics, outperforming AV-GridNet$_n$ by 0.05 in PESQ and 0.1 dB in SI-SDR. For the WER evaluation, we followed the methodology described in Richter et al. (2023b) and used the NeMo ASR toolkit (NVIDIA, 2021) to generate transcripts for evaluation. AV-GridNet$_n$ and AV-Gen deliver comparable performance to our system in terms of WER. To summarize, the evaluation and comparison results on multiple datasets demonstrate the strong performance and robustness of our proposed system, especially in very low SNR conditions.

4.4. Ablation studies

4.4.1. Two-stage fusion strategy

To examine the performance gain of the proposed two-stage fusion strategy, we conduct experiments on the COG-MHEAR AVSE challenge dataset and the LRS3-AudioSet dataset. As shown in Table 4, our system with signal-level fusion and embedding-level fusion achieves nearly identical performances, slightly outperforming AV-CrossNet which uses pre-extracted, fixed visual embeddings. However, the proposed two-stage fusion strategy provides a PESQ improvement over each individual fusion strategy. The ablation results on the LRS3-AudioSet dataset are shown in Table 5. On this test set, signal-level fusion and embedding-level

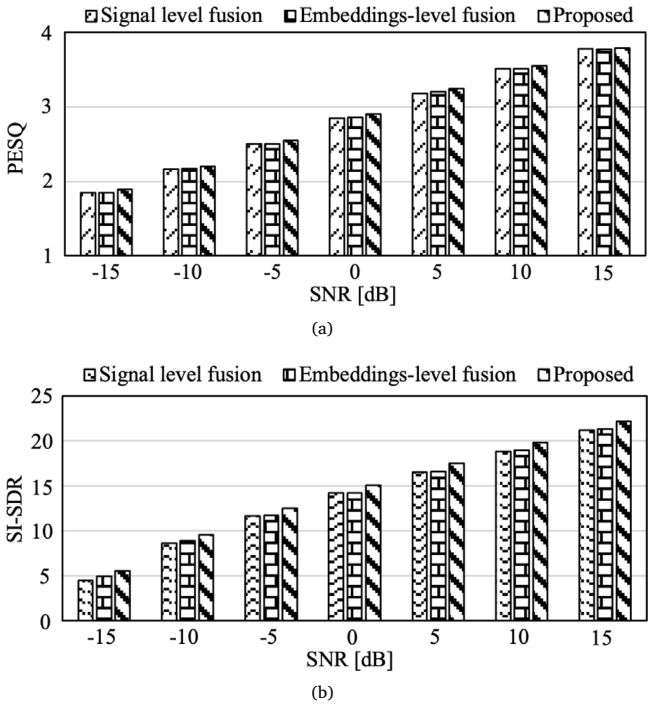

Fig. 4. PESQ and SI-SDR scores across different SNR levels on LRS3-AudioSet. Results are shown for signal-level fusion, embedding-level fusion, and the proposed method.

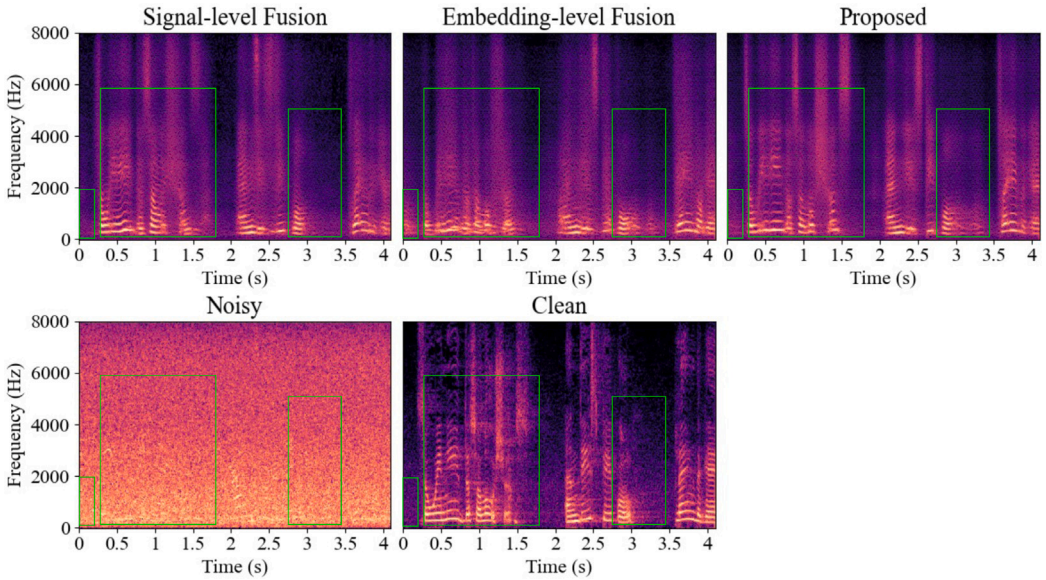

Fig. 5. Spectrograms of enhanced, noisy, and clean utterances. Green boxes highlight the different results of signal-level, embedding-level, and two-stage fusion.

fusion achieve close results, and our proposed system benefits from the two-stage fusion strategy. Performance improvements over the audio-only baseline of TF-CrossNet align with those reported in Table 4, confirming the effectiveness of our approach.

To further analyze the potentially complementary effects of the two fusion strategies, Fig. 4 presents PESQ and SI-SDR scores across different SNR levels on the LRS3-AudioSet dataset. While signal-level fusion and embedding-level fusion achieve similar performance, the proposed two-stage fusion consistently outperforms both across all SNR conditions. To gain a deeper insight, Fig.

Table 6
Ablation results of different model configurations on the LRS3-AudioSet.

Method	AV	Params (M)	B	F	PESQ↑	STOI↑	SI-SDR↑
Unprocessed	–	–	–	–	1.32	0.72	0.08
TF-CrossNet-small	✗	4.3	6	129	2.55	0.88	12.81
Proposed-small	✓	5.7	6	129	2.66	0.90	12.64
TF-CrossNet-wide	✗	6.2	8	321	2.70	0.89	13.69
Proposed-wide	✓	9.2	8	321	2.82	0.92	13.46
TF-CrossNet	✗	7.3	12	129	2.77	0.90	13.90
Proposed	✓	9.6	12	129	**2.87**	**0.93**	**13.93**

5 shows the spectrograms of noisy and enhanced input signals. The noise is beach babble noise and the SNR is −10 dB. Green boxes highlight the regions where the single-stage fusion models exhibit relatively distinct results. The two-stage fusion appears to provide more balanced enhancement.

4.4.2. System configuration

We now examine the impact of different system configurations (see Fig. 1) on the performance of our proposed system. We evaluate three configurations for both the audio-only TF-CrossNet and the proposed system: small configuration with $B = 6$ and $F = 129$ (see Section 2.B), wide configurations with $B = 8$ and $F = 321$, and full-sized configuration $B = 12$ and $F = 129$, which yields the best performance. The small configuration is the smallest model with approximately half the size of the full-sized setup. The wide setup has a larger convolutional kernel but half the depth relative to the full-sized configuration, resulting in a slightly smaller size. This wide configuration is chosen to match the time–frequency resolution of the LipVoicer Mel spectrogram feature, aligning complex spectrograms and Mel spectrograms in the signal-level fusion stage. As shown in Table 6, SE performance correlates with model size, with PESQ scores of 2.66, 2.82, and 2.87 for small, wide, and full-sized configuration, respectively.

5. Conclusion

To conclude, we have proposed an audiovisual speech enhancement algorithm that also addresses voice activity detection. The proposed algorithm is evaluated on multiple AV datasets, and achieves the state-of-the-art speech enhancement performance, including VAD and ASR tasks. The two-stage audiovisual fusion leverages generative features using LipVoicer and regressive visual embeddings using a ResNet18 encoder to enhance speech quality and improve noise robustness. Looking ahead, there are several promising directions for future research. Exploring more advanced visual encoders could lead to further performance improvements. Developing real-time implementation and compressing model size for deployment presents exciting opportunities for practical applications, such as AVSE in mobile devices or assistive technologies.

CRediT authorship contribution statement

Cheng Yu: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Vahid Ahmadi Kalkhorani:** Writing – review & editing, Validation, Methodology, Data curation, Conceptualization. **Buye Xu:** Writing – review & editing, Supervision, Project administration, Conceptualization. **DeLiang Wang:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: DeLiang Wang reports financial support was provided by Meta. DeLiang Wang reports equipment, drugs, or supplies was provided by Ohio Supercomputer Center. DeLiang Wang reports equipment, drugs, or supplies was provided by NCSA Delta Supercomputer Center. DeLiang Wang reports equipment, drugs, or supplies was provided by Pittsburgh Supercomputing Center. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by a Meta contract to Ohio State University, United States, the Ohio Supercomputer Center, the NCSA Delta Supercomputer Center (OCI 2005572), and the Pittsburgh Supercomputer Center (NSF ACI-1928147).

Data availability

The authors do not have permission to share data.

References

Afouras, T., Chung, J.S., Senior, A., Vinyals, O., Zisserman, A., 2018a. Deep audio-visual speech recognition. IEEE Trans. Pattern Anal. Mach. Intell. 44, 8717–8727.

Afouras, T., Chung, J.S., Zisserman, A., 2018b. The conversation: Deep audio-visual speech enhancement. In: Proc. INTERSPEECH. pp. 3244–3248.

Afouras, T., Chung, J.S., Zisserman, A., 2018c. LRS3-TED: a large-scale dataset for visual speech recognition. arXiv:1809.00496.

Barker, J., Marxer, R., Vincent, E., Watanabe, S., 2015. The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines. In: Proc. ASRU. pp. 504–511.

Blanco, A.L.A., Valentini-Botinhao, C., Klejch, O., Gogate, M., Dashtipour, K., Hussain, A., Bell, P., 2023. AVSE Challenge: Audio-visual speech enhancement challenge. In: Proc. SLT. pp. 465–471.

Chen, H., Mira, R., Petridis, S., Pantic, M., 2024. RT-LA-VocE: Real-time low-SNR audio-visual speech enhancement. In: Proc. INTERSPEECH. pp. 2215–2219.

Defossez, A., Synnaeve, G., Adi, Y., 2020. Real time speech enhancement in the waveform domain. In: Proc. INTERSPEECH. pp. 3291–3295.

Fonseca, E., Pons Puig, J., Favory, X., Font Corbera, F., Bogdanov, D., Ferraro, A., Oramas, S., Porter, A., Serra, X., 2017. Freesound datasets: a platform for the creation of open audio datasets. In: Proc. ISMIR. International Society for Music Information Retrieval, pp. 486–493.

Gabbay, A., Ephrat, A., Halperin, T., Peleg, S., 2018. Seeing through noise: Visually driven speaker separation and enhancement. In: Proc. ICASSP. pp. 3051–3055.

Gao, R., Grauman, K., 2021. VisualVoice: Audio-visual speech separation with cross-modal consistency. In: Proc. CVPR. pp. 15490–15500.

Graetzer, S., Barker, J., Cox, T.J., Akeroyd, M., Culling, J.F., Naylor, G., Porter, E., Viveros Munoz, R., 2021. Clarity-2021 challenges: Machine learning challenges for advancing hearing aid processing. In: Proc. INTERSPEECH. pp. 686–690.

Gulati, A., Qin, J., Chiu, o., 2020. Conformer: Convolution-augmented transformer for speech recognition. In: Proc. INTERSPEECH. pp. 5036–5040.

Hao, X., Su, X., Wen, S., Wang, Z., Pan, Y., Bao, F., Chen, W., 2020. Masking and inpainting: A two-stage speech enhancement approach for low SNR and non-stationary noise. In: Proc. ICASSP. pp. 6959–6963.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proc. ICCV. pp. 1026–1034.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proc. CVPR. pp. 770–778.

Hendrycks, D., Gimpel, K., 2016. Gaussian error linear units (GeLUs). arXiv:1606.08415.

Hou, J.-C., Wang, S.-S., Lai, Y.-H., Tsao, Y., Chang, H.-W., Wang, H.-M., 2018. Audio-visual speech enhancement using multimodal deep convolutional neural networks. IEEE Trans. Emerg. Top. Comput. Intell. 2, 117–128.

Hsu, W.-N., Remez, T., Shi, B., Donley, J., Adi, Y., 2023. ReVISE: Self-supervised speech resynthesis with visual input for universal and generalized speech regeneration. In: Proc. CVPR. pp. 18795–18805.

Hu, Y., Loizou, P.C., 2008. Evaluation of objective quality measures for speech enhancement. IEEE Trans. Audio, Speech, Lang. Process. 16, 229–238.

Jensen, J., Taal, C.H., 2016. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. IEEE/ACM Trans. Audio, Speech, Lang. Process. 24, 2009–2022.

Jung, C., Lee, S., Kim, J.-H., Chung, J.S., 2024. FlowAVSE: Efficient audio-visual speech enhancement with conditional flow matching. In: Proc. INTERSPEECH. pp. 2210–2214.

Kalkhorani, V.A., Kumar, A., Tan, K., Xu, B., Wang, D.L., 2024. Audiovisual speaker separation with full-and sub-band modeling in the time-frequency domain. In: Proc. ICASSP. pp. 12001–12005.

Kalkhorani, V.A., Wang, D.L., 2024. TF-CrossNet: Leveraging global, cross-band, narrow-band, and positional encoding for single-and multi-channel speaker separation. IEEE/ACM Trans. Audio, Speech, Lang. Process. 32, 4999–5009, Code available: https://github.com/ahmadikalkhorani/CrossNet.

Kalkhorani, V.A., Yu, C., Kumar, A., Tan, K., Xu, B., Wang, D.L., 2025. AV-CrossNet: an audiovisual complex spectral mapping network for speech separation by leveraging narrow-and cross-band modeling. IEEE J. Sel. Top. Signal Process. 19, 685–694.

Kim, D., Chung, S.-W., Han, H., Ji, Y., Kang, H.-G., 2023. HD-Demucs: General speech restoration with heterogeneous decoders. In: Proc. INTERSPEECH. pp. 3829–3833.

Kim, G., Lu, Y., Hu, Y., Loizou, P.C., 2009. An algorithm that improves speech intelligibility in noise for normal-hearing listeners. J. Acoust. Soc. Am. 126, 1486–1494.

Lea, C., Flynn, M.D., Vidal, R., Reiter, A., Hager, G.D., 2017. Temporal convolutional networks for action segmentation and detection. In: Proc. CVPR. pp. 156–165.

Liao, S., Lan, S., Zachariah, A.G., 2024. EVA-GAN: Enhanced various audio generation via scalable generative adversarial networks. arXiv:2402.00892.

Loizou, P.C., 2013. Speech Enhancement: Theory and Practice. CRC Press.

Ma, P., Petridis, S., Pantic, M., 2022. Visual speech recognition for multiple languages in the wild. Nat. Mach. Intell. 4, 930–939.

Michelsanti, D., Tan, Z.-H., Zhang, S.-X., Xu, Y., Yu, M., Yu, D., Jensen, J., 2021. An overview of deep-learning-based audio-visual speech enhancement and separation. IEEE/ACM Trans. Audio, Speech, Lang. Process. 29, 1368–1396.

Mira, R., Xu, B., Donley, J., Kumar, A., Petridis, S., Ithapu, V.K., Pantic, M., 2023. LA-VocE: Low-SNR audio-visual speech enhancement using neural vocoders. In: Proc. ICASSP. pp. 1–5.

Morrone, G., Bergamaschi, S., Pasa, L., Fadiga, L., Tikhanoff, V., Badino, L., 2019. Face landmark-based speaker-independent audio-visual speech enhancement in multi-talker environments. In: Proc. ICASSP. pp. 6900–6904.

Nonaka, Y., Leow, C.S., Kobayashi, A., Utsuro, T., Nishizaki, H., 2021. Voice activity detection for live speech of baseball game based on tandem connection with speech/noise separation model. In: Proc. INTERSPEECH. pp. 351–355.

NVIDIA, 2021. NeMo: A toolkit for building AI-powered conversational applications. URL https://developer.nvidia.com/nvidia-nemo.

Pan, Z., Ge, M., Li, H., 2022a. A hybrid continuity loss to reduce over-suppression for time-domain target speaker extraction. In: Proc. INTERSPEECH. pp. 1786–1790.

Pan, Z., Ge, M., Li, H., 2022b. USEV: Universal speaker extraction with visual cue. IEEE/ACM Trans. Audio, Speech, Lang. Process. 30, 3032–3045.

Pan, Z., Wichern, G., Masuyama, Y., Germain, F.G., Khurana, S., Hori, C., Le Roux, J., 2023. Scenario-aware audio-visual TF-GridNet for target speech extraction. In: Proc. ASRU. pp. 1–8.

Prajwal, K.R., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C., 2020a. Learning individual speaking styles for accurate lip to speech synthesis. In: Proc. CVPR. pp. 13796–13805.

Prajwal, K., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C., 2020b. A lip sync expert is all you need for speech to lip generation in the wild. In: Proc. Multimedia. pp. 484–492.

Radford, A., et al., 2022. Whisper: Robust speech recognition via large-scale self-supervised learning. URL https://github.com/openai/whisper.

Reddy, C.K., Gopal, V., Cutler, R., et al., 2021. The INTERSPEECH 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results. In: Proc. INTERSPEECH. pp. 2492–2496.

Richter, J., Frintrop, S., Gerkmann, T., 2023a. Audio-visual speech enhancement with score-based generative models. In: Proc. ITG Speech Communication. VDE, pp. 275–279.

Richter, J., Welker, S., Lemercier, J.-M., Lay, B., Gerkmann, T., 2023b. Speech enhancement and dereverberation with diffusion-based generative models. IEEE/ACM Trans. Audio, Speech, Lang. Process. 31, 2351–2364.

Rouditchenko, A., Gong, Y., Thomas, S., Karlinsky, L., Kuehne, H., Feris, R., Glass, J., 2024. Whisper-flamingo: Integrating visual features into Whisper for audio-visual speech recognition and translation. In: Proc. INTERSPEECH. pp. 2420–2424.
Sadeghi, M., Alameda-Pineda, X., 2021. Mixture of inference networks for VAE-based audio-visual speech enhancement. IEEE Trans. Signal Process. 69, 1899–1909.
Sadeghi, M., Leglaive, S., Alameda-Pineda, X., Girin, L., Horaud, R., 2020. Audio-visual speech enhancement using conditional variational auto-encoders. IEEE/ACM Trans. Audio, Speech, Lang. Process. 28, 1788–1800.
Schwartz, J.-L., Berthommier, F., Savariaux, C., 2004. Seeing to hear better: evidence for early audio-visual interactions in speech identification. Cognition 93, B69–B78.
Shahid, M., Beyan, C., Murino, V., 2021. S-VVAD: Visual voice activity detection by motion segmentation. In: Proc. Winter Conference on Applications of Computer Vision. pp. 2332–2341.
Sumby, W.H., Pollack, I., 1954. Visual contribution to speech intelligibility in noise. J. Acoust. Soc. Am. 26, 212–215.
Sun, Z., Wang, Y., Cao, L., 2020. An attention based speaker-independent audio-visual deep learning model for speech enhancement. In: Proc. MMM. Springer, pp. 722–728.
Taal, C.H., Hendriks, R.C., Heusdens, R., Jensen, J., 2010. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In: Proc. ICASSP. pp. 4214–4217.
Tan, Z.-H., Sarkar, A., Dehak, N., 2020. rVAD: An unsupervised segment-based robust voice activity detection method. Comput. Speech & Lang. 59, 1–21.
Tan, X., Zhang, X.-L., 2021. Speech enhancement aided end-to-end multi-task learning for voice activity detection. In: Proc. ICASSP. pp. 6823–6827.
Tang, M., Huang, H., Zhang, W., He, L., 2024. Phase continuity-aware self-attentive recurrent network with adaptive feature selection for robust VAD. In: Proc. ICASSP. pp. 11506–11510.
Tesema, F.B., Gu, J., Song, W., Wu, H., Zhu, S., Lin, Z., 2023. Efficient audiovisual fusion for active speaker detection. IEEE Access 11, 45140–45153.
Thiemann, J., Ito, N., Vincent, E., 2013. DEMAND: A collection of multi-channel recordings of acoustic noise in diverse environments. Meet. Acoust. 1–6.
Wang, D.L., Chen, J., 2018. Supervised speech separation based on deep learning: An overview. IEEE/ACM Trans. Audio, Speech, Lang. Process. 26, 1702–1726.
Wang, Z.-Q., Cornell, S., Choi, S., Lee, Y., Kim, B.-Y., Watanabe, S., 2023. TF-GridNet: Integrating full-and sub-band modeling for speech separation. IEEE/ACM Trans. Audio, Speech, Lang. Process. 3221–3236.
Wang, H., Wang, D.L., 2023. Cross-domain diffusion based speech enhancement for very noisy speech. In: Proc. ICASSP. pp. 1–5.
Xu, X., Wang, Y., Xu, D., Peng, Y., Zhang, C., Jia, J., Chen, B., 2022. VSEGAN: Visual speech enhancement generative adversarial network. In: Proc. ICASSP. pp. 7308–7311.
Yang, K., Marković, D., Krenn, S., Agrawal, V., Richard, A., 2022. Audio-visual speech codecs: Rethinking audio-visual speech enhancement by re-synthesis. In: Proc. CVPR. pp. 8227–8237.
Yemini, Y., Shamsian, A., Bracha, L., Gannot, S., Fetaya, E., 2024. LipVoicer: Generating speech from silent videos guided by lip reading. In: Proc. ICLR. pp. 1–20.
Zhao, Y., Champagne, B., 2022. An efficient transformer-based model for voice activity detection. In: Proc. MLSP Workshop. pp. 1–6.
Zhu, Z., Yang, H., Tang, M., Yang, Z., Eskimez, S.E., Wang, H., 2023. Real-time audio-visual end-to-end speech enhancement. In: Proc. ICASSP. pp. 1–5.
Zhuang, Y., Tong, S., Yin, M., Qian, Y., Yu, K., 2016. Multi-task joint-learning for robust voice activity detection. In: Proc. ISCSLP. pp. 1–5.