BAV-MossFormer2: Enhanced MossFormer2 for Binaural Audio-Visual Speech Enhancement

Wenze Ren[1], *Kai Li*[2], *Rong Chao*[1], *Junjie Li*[3], *Zilong Huang*[3], *Shafique Ahmed*[4], *You-Jin Li*[1], *Kuo-Hsuan Hung*[1], *Syu-Siang Wang*[5], *Hsin-Min Wang*[4], *Yu Tsao*[4]

[1]National Taiwan University,[2]Tsinghua University,[3]The Hong Kong Polytechnic University,[4]Academia Sinica,[5]Yuanze University

Abstract

This paper presents a novel audio-visual speech enhancement architecture for binaural audio channels, the BAV-MossFormer2. The framework is based on an enhanced audio encoder with cross-attention mechanisms, enabling complex left-right channel interaction and fusion. It also adopts adaptive dynamic modules to utilise multi-scale modal features and learnable attention weights, thereby optimising the fusion of audio-visual representations. Finally, the advanced MossFormer2 architecture is employed to achieve effective speech enhancement. On the COG-MHEAR Audio-Visual Speech Enhancement Challenge 4, our proposed BAV-MossFormer2 architecture not only outperforms baseline methods but also significantly improves speech quality and intelligibility metrics under various noise conditions. These results underscore the significance of our proposed binaural interaction strategy and adaptive fusion method in achieving robust binaural audio-visual speech enhancement.

Index Terms: speech enhancement, audio-visual, multimodal fusion, mossformer2

1. Introduction and Related Work

Speech enhancement in noisy settings remains a central challenge in audio processing, with practical applications that range from hearing aids to video conferencing. Traditional, audio-only methods have advanced the field, yet they struggle when background noise becomes severe. By contrast, audio-visual approaches can exploit the complementary strengths of each modality: while sound is easily masked, lip-movement cues remain stable even under heavy acoustic interference. Integrating these visual signals, therefore, provides a noise-resistant pathway for restoring clean speech in the most demanding conditions.

Recent advances in audio-visual speech enhancement(AVSE) leverage deep learning to uncover cross-modal correlations. Early work concatenated audio and visual features and applied convolutional neural networks [1] or U-net architectures [2] for speech enhancement. During this period, researchers explored various audio and visual feature fusion strategies [3, 4, 5, 6], and some approaches first performed speech separation before selecting the target speaker based on visual features [7]. Furthermore, audio-visual speech enhancement has driven the development of visual front-end models, where pre-trained ResNet-based models can extract robust visual features from the mouth region. However, most existing AVSE systems suffer from several key limitations. First, they primarily focus on monaural audio processing, neglecting the valuable spatial information present in stereo signals. Second, most current fusion strategies employ simple concatenation or basic attention mechanisms, lacking the adaptive capability for dynamic cross-modal integration. This highlights the urgent need for further research and development in this area to create more effective and versatile AVSE systems.

Therefore, this paper proposes an innovative architecture, BAV-MossFormer2, based on MossFormer2. We propose an enhanced stereo audio encoder that uses an attention mechanism between the left and right channels to achieve channel interaction. At the same time, we propose an adaptive dynamic fusion module that utilizes learnable attention weights to process multi-scale audio and visual features, thereby achieving context-aware audio-video feature fusion through a gating mechanism. Experimental results demonstrate that BAV-MossFormer2 outperforms the official baseline provided by AVSE4.

2. Proposed Method

As shown in the figure 1, the proposed BAV-MossFormer2 architecture mainly consists of three main stages: visual feature extraction, enhanced stereo audio encoding, and cross-modal fusion. The model input is dual-channel noisy speech, and the output is dual-channel speech after noise reduction.

2.1. Visual Phase

The visual processing pipeline extracts robust speaker facial features from the input video sequence. We adopt a pre-trained visual front-end and combine it with temporal convolution to further extract visual features during speech.

We utilize a pre-trained model comprising 3D convolution blocks and an 18-layer ResNet architecture to capture spatiotemporal speaker facial dynamics. The input video frames are first transposed to adapt to the model's dimensions. To further capture the temporal dependency of speaker facial motion, we employ a visual temporal convolution network consisting of five Conv1D blocks, each with residual connections.

By using pre-trained parameters from a large-scale lip-reading dataset and freezing the visual front-end weights, we ensure robust visual feature extraction while adapting the TCN layers to the speech enhancement task.

2.2. Audio Phase

The audio stage processes stereo input through an enhanced encoder. This design utilises cross-attention and self-attention mechanisms to capture inter-channel correlations while preserving spatial audio features.

The input stereo audio signal is first decomposed into left and right channels and processed independently through separate 1D convolutions. To achieve information interaction be-

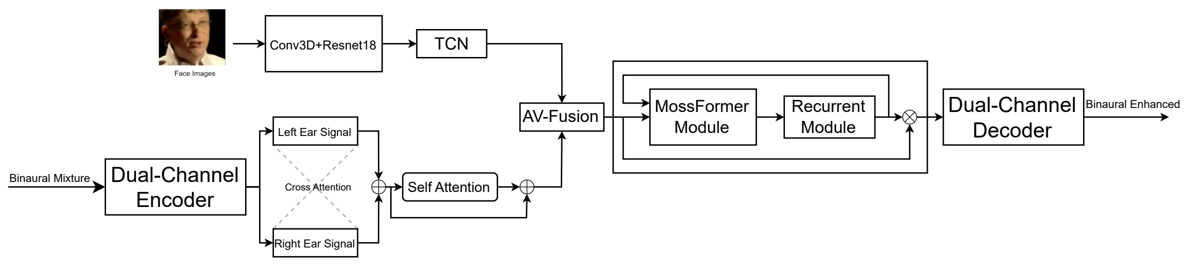

Figure 1: *The overall architecture of the proposed audio-visual speech enhancement. The AV-Fusion module is the proposed adaptive dynamic fusion module.*

tween stereo channels, we designed a symmetric cross-attention mechanism. The left-channel features are used as query vectors, and the right-channel features are used as key-value pairs for attention calculations; the same applies in reverse. This bidirectional cross-attention design enables both left and right channels to obtain complementary information from the other channel, thereby enhancing the understanding of the spatial audio scene. All cross-attention outputs are connected to the original inputs via residual connections and processed through layer normalisation, ensuring the preservation of the original characteristics of the stereo input while achieving information fusion between the two channels.

The enhanced left and right channel features are concatenated and fused in the feature dimension. Finally, the fused features are modelled for global temporal dependencies via self-attention, capturing long-range temporal dependencies, and outputting the final stereo-enhanced audio encoding representation.

2.3. Fusion Phase

The fusion stage integrates audio and visual features through our novel adaptive dynamic fusion module, which employs multi-scale processing and context-aware mechanisms.

Audio and visual features are processed through parallel convolutional layers with different kernel sizes to capture multi-scale audio and visual features. Then, we use global features from all scales to calculate context-aware weights and generate standardised attention weights for each scale.

Scale-weighted features are processed through cross-modal attention to enhance intermodal relevance, while a gating mechanism is used to control the contribution of the audio and visual modalities. Finally, the features from both modalities are connected and processed through a fusion layer, whose output contains audio features with residual connections.

3. Experiment and Result

We utilized the official script provided by the AVSE4 Challenge to generate the data, with the training set comprising 34,524 scenes and the development set comprising 3,306 scenes. The target data comes from the LRS3 dataset, which has three types of interference sources: competing speakers, non-speech noise, and music. The script will directly create binaural signals.

Training process The visual pre-trained model is fine-tuned together, with a batch size set to 1, a learning rate set to 1e-4, and the Adam optimiser used. Training is stopped if the validation loss does not improve within six epochs.

The table 1 shows our results, which demonstrate that BAV-MossFormer2 significantly outperforms the baseline.

Table 1: *Performance on the AVSEC-4 development set (Binaural audio).*

Method	MBSTOI	PESQ
Noisy Input	0.4161	1.30
Baseline (Binaural)	-	1.21
Ours (BAV-MossFormer2)	**0.687**	**1.86**

4. Conclusion

This paper introduces BAV-MossFormer2, an innovative architecture for audio-visual speech enhancement in binaural channels, which addresses the key limitations of existing AVSE systems. We introduce an enhanced stereo audio encoder that can effectively achieve inter-channel information exchange. The adaptive dynamic fusion module we propose achieves significant improvements over traditional fusion methods. Our experimental results on the AVSE4 challenge dataset demonstrate that BAV-MossFormer2 significantly outperforms the official baseline, thereby setting a new benchmark for future research in binaural audio-visual speech enhancement and inspiring further advancements in the field.

5. References

[1] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, and H.-M. Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 117–128, 2018.

[2] S. Ahmed, C.-W. Chen, W. Ren, C.-J. Li, E. Chu, J.-C. Chen, A. Hussain, H.-M. Wang, Y. Tsao, and J.-C. Hou, "Deep complex u-net with conformer for audio-visual speech enhancement," 2023.

[3] K. Li, R. Yang, F. Sun, and X. Hu, "Iianet: An intra- and inter-modality attention network for audio-visual speech separation," 2024. [Online]. Available: https://arxiv.org/abs/2308.08143

[4] S. Pegg, K. Li, and X. Hu, "Rtfs-net: Recurrent time-frequency modelling for efficient audio-visual speech separation," 2024. [Online]. Available: https://arxiv.org/abs/2309.17189

[5] W. Sang, K. Li, R. Yang, J. Huang, and X. Hu, "A fast and lightweight model for causal audio-visual speech separation," 2025. [Online]. Available: https://arxiv.org/abs/2506.06689

[6] I.-C. Chern, K.-H. Hung, Y.-T. Chen, T. Hussain, M. Gogate, A. Hussain, Y. Tsao, and J.-C. Hou, "Audio-visual speech enhancement and separation by utilizing multi-modal self-supervised embeddings," in *2023 ICASSPW*, 2023, pp. 1–5.

[7] W. Ren, K.-H. Hung, R. Chao, Y. Li, H.-M. Wang, and Y. Tsao, "Robust audio-visual speech enhancement: Correcting misassignments in complex environments with advanced post-processing," in *2024 27th O-COCOSDA*, 2024, pp. 1–6.