Audio-Visual Speech Enhancement in Complex Scenarios with Separation and Dereverberation Joint Modeling

Zhan Jin[1,2], Jiarong Du[3,4], Peijun Yang[3,4], Juan Liu[3,4], and Ming Li[2,3✉⋆]

[1] School of Computer Science, Wuhan University, Wuhan, China
[2] School of Artificial Intelligence, Chinese University of Hong Kong, Shenzhen, China
[3] School of Artificial Intelligence, Wuhan University, Wuhan, China
[4] School of Cyber Science and Engineering, Wuhan University, Wuhan, China
mingli369@cuhk.edu.cn

Abstract. Audio-visual speech enhancement (AVSE) uses visual cues to extract a target speaker from acoustically complex mixtures containing multiple interferers and reverberation. We present a system-level separation–dereverberation framework for this setting. The submitted AVSEC-4 system consists of a binaural TFGridNet-based separation model followed by SGMSE+ dereverberation, while progressive layer-wise supervision is applied to the separation stage. We additionally study a binaural-input, monaural-output separation model for complete-test-set analysis and a SkipConvNet-based configuration for joint optimization. Under the official AVSEC-4 protocol, the submitted system obtains competitive objective results and a high word-accuracy score in the subjective listening evaluation. Demo samples are available at https://helpful-rabanadas-47fa43.netlify.app/.

Keywords: Audio-Visual Speech Enhancement · Target Speaker Extraction · Dereverberation

1 Introduction

Speech signals in real-world environments are frequently corrupted by interfering sounds, background noise, and reverberation, posing significant challenges to human perception and downstream applications such as automatic speech recognition [1]. Separating a target speaker's speech from such mixtures, known as the "cocktail party problem" [2], remains a critical yet difficult task. To address this, blind source separation (BSS) methods have been extensively studied [3–5]. However, they typically require prior knowledge of the number of active speakers and, more critically, lack an intrinsic mechanism to associate the separated streams with specific speaker identities. Target Speaker Extraction (TSE) offers a more directed solution by conditioning the extraction process on a reference

⋆ Corresponding author: Ming Li

cue that identifies the target speaker. This enables selective isolation of a target speaker without constraints on the total number of speakers. Early TSE systems typically rely on a pre-recorded audio reference [6–9]. However, obtaining such a sample in advance is often impractical. This key limitation has driven the exploration of alternative modalities for speaker conditioning.

Consequently, visual cues, particularly lip movements, have been used as a synchronous auxiliary modality to condition the extraction process, forming the basis of AVSE systems [10–14]. Pre-trained lip-reading networks have also been adopted to extract speech-related visual representations and fuse them with audio for speaker extraction [15–17]. Subsequent studies investigated more elaborate fusion strategies for combining visual cues with the mixed audio spectrogram [18–20]. Furthermore, neurocognition-inspired fusion methods have been reported to improve AVSE performance in recent studies [21, 22]. Nevertheless, the behavior of AVSE systems under complex acoustic conditions, including reverberation and diverse interferers, remains an important evaluation issue. This setting poses a dual challenge: models trained on relatively simple mixtures may not generalize well to complex acoustic conditions, while training directly on heavily corrupted data can make it difficult to obtain clean speech estimates.

To address these limitations, we introduce a set of system-level modifications to the baseline framework established in [23], with the goal of improving speech extraction in complex acoustic environments. First, we use a dynamic data-mixing strategy to generate training mixtures with controlled acoustic difficulty, thereby preparing the data for the subsequent progressive training scheme. Second, we adopt a two-stage processing pipeline that performs target-speaker extraction before dereverberation. In the first stage, a progressive loss training scheme [24] is incorporated into the separation network so that intermediate layers are supervised with targets of gradually increasing SNR. The submitted challenge system uses a binaural separation model followed by SGMSE+ dereverberation. For additional complete-test-set analysis, we compare binaural-output and monaural-output separation models that both receive binaural mixtures, and we further examine a SkipConvNet-based jointly trained configuration. The contribution of this work is therefore positioned as the adaptation and empirical validation of a separation–dereverberation training recipe for the AVSEC-4 complex AVSE setting, rather than the introduction of a new backbone architecture. Under the official AVSEC-4 evaluation protocol, the submitted system obtains competitive scores in Perceptual Evaluation of Speech Quality (PESQ) [25], Short-Time Objective Intelligibility (STOI) [26], and Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) [27], and it achieves a high score in the subjective listening evaluation.

2 Methodology

2.1 Overall Model Architecture

Our architecture builds upon TFGridNet [5], a time-frequency-domain separation network based on GridBlock modules. Its extension to audio-visual tasks,

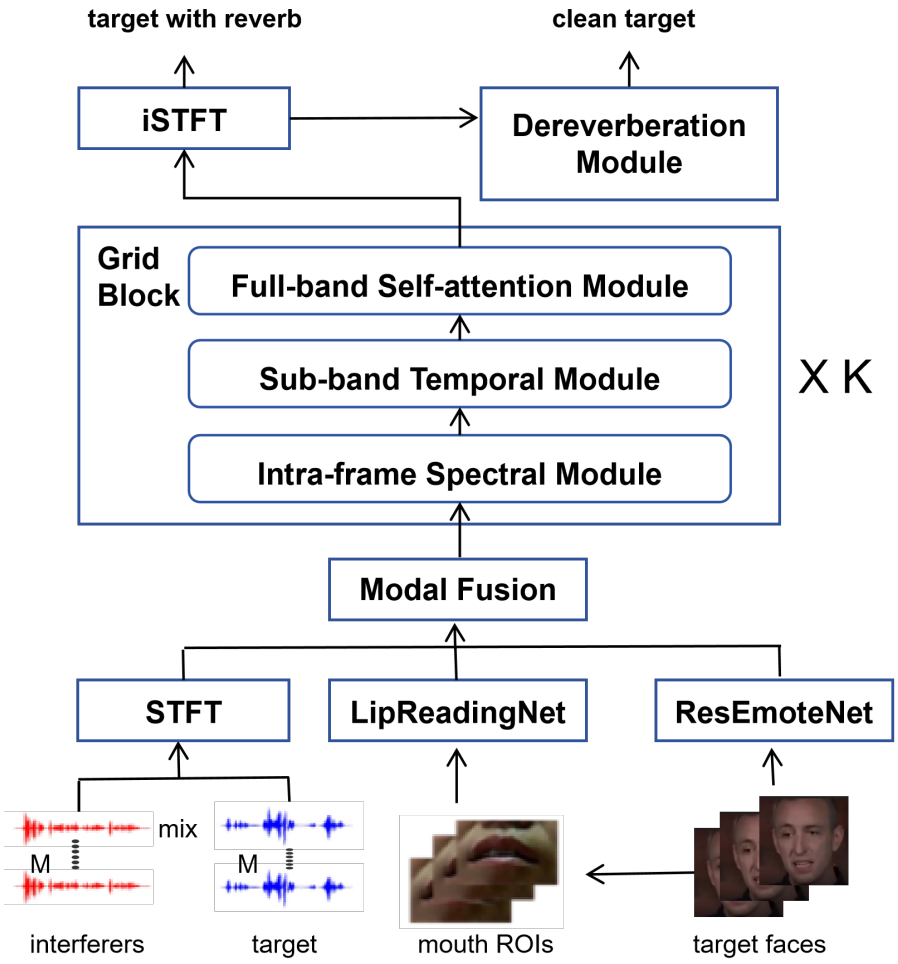


Fig. 1: The proposed model architecture, which includes visual feature extractors (for lip and expression), an audio-visual fusion module, the TFGridNet-based separation network and a dereverberation module.

AV-GridNet [28], showed that visual cues can be incorporated into this framework. Following this direction, recent work has further simplified the fusion process by combining audio and visual features before separation [23]. In addition to lip movements, we explore facial expression representations as an auxiliary visual cue, following the multi-enrollment setting considered in [29]. Motivated by these studies, we construct a multi-cue audio-visual fusion network.

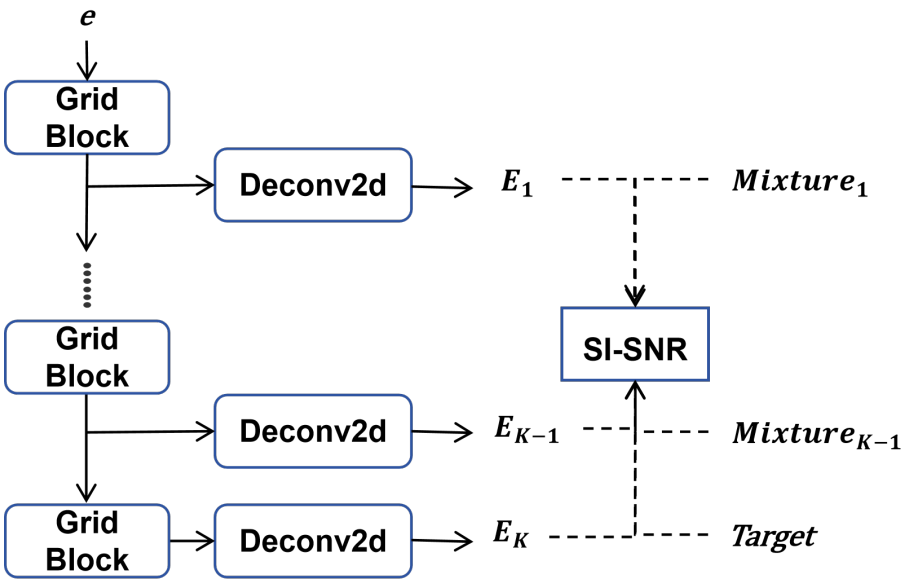


Fig. 2: Progressive training. In this work, K is set to 6. E_k represents the intermediate estimate output by the k-th separation module layer. $\mathrm{Mixture}_k$ denotes the intermediate target assigned to the k-th layer. The intermediate targets are generated by increasing the SNR in fixed 5 dB steps from the initial mixture toward the final reverberant target speech. Target refers to the reverberant target speaker speech used as the final separation target.

The proposed model, illustrated in Figure 1, follows an encode-fuse-separate paradigm. First, multi-modal features are extracted in parallel. The audio stream is transformed into a time-frequency representation via the Short-Time Fourier Transform (STFT), which serves as the acoustic input. For the visual stream, we use two types of cues. Lip movement features are encoded from consecutive video frames using a pre-trained lip-reading network [15]. Facial expression features are extracted using an expression estimation model [30] and are used as an additional visual representation. These visual features are temporally interpolated to match the audio feature rate. Subsequently, the audio spectrogram and the two visual feature maps are concatenated along the channel dimension at each time step. The fused feature map serves as the input to the core separation module. We

adopt the TFGridNet [5] architecture, which is built upon a series of GridBlocks. Each GridBlock performs interleaved intra-band and cross-band modeling in the time-frequency domain to capture local spectral structures and long-range temporal dependencies. The integrated visual cues provide target-speaker-related guidance for selecting time-frequency components associated with the visible speaker.

2.2 Separation Before Dereverberation and Progressive Training

Previous research indicates that the reverberation characteristics of training data can influence model performance [31]. When anechoic clean speech is set as the training target for reverberant mixtures, the acoustic mismatch may increase optimization difficulty and degrade extraction quality. Conversely, using reverberant speech as the target can ease the separation task but leaves residual reverberation in the output [31]. Motivated by this trade-off, we adopt a two-stage paradigm, namely separation before dereverberation, to decouple the two objectives. This decoupling allows the separation stage to focus on isolating the target speaker's reverberant speech, while the subsequent dereverberation stage removes residual reverberation from the separated signal.

To stabilize the optimization of the separation network under challenging acoustic conditions, we adopt a progressive loss training strategy [24]. We construct each training sample by mixing the reverberant target speech with interfering speech at a random Signal-to-Noise Ratio (SNR), uniformly sampled between -18 dB and 6 dB. From this original mixture, we generate a sequence of $K-1$ intermediate targets with gradually increasing SNR. The SNR of adjacent intermediate targets is separated by a fixed step of 5 dB, and the last target is the reverberant target speech itself. This design forms a curriculum from heavily interfered mixtures to the final reverberant target. As illustrated in Figure 2, each GridBlock is followed by a dedicated Deconv2d layer that produces an intermediate waveform estimate for training supervision. A scale-invariant signal-to-distortion ratio (SI-SDR) loss [27] is computed between each estimate and its assigned intermediate target. The total separation loss is the mean of all K layer-wise losses, which is defined as:

$$\mathcal{L}_{\text{total}} = \frac{1}{K} \sum_{k=1}^{K} \mathcal{L}^{(k)}$$

where $K = 6$ denotes the number of GridBlocks, $\mathcal{L}^{(k)}$ represents the loss at the k-th layer, and $\mathcal{L}_{\text{total}}$ denotes the total loss of the separation model. This progressive curriculum provides layer-wise supervision and encourages the network to move from coarse interference suppression to finer target-speech estimation in deeper layers.

The reverberant speech output from the first stage is then fed into the second, dereverberation stage. In the cascaded version of the system, we use a pre-trained dereverberation model [32] to reduce residual reverberation and obtain the final speech estimate. Specifically, we use SGMSE+ [32], a score-based generative

diffusion model that operates on the complex-valued STFT domain. This representation allows the model to process both magnitude and phase information. The model iteratively refines the speech estimate through a generative diffusion process and uses 30 diffusion steps during inference in our setting. Since this module is used as a post-processing component in the cascaded system, the separated speech can be dereverberated without changing the separation backbone.

2.3 Joint Training of the Separation-Dereverberation Pipeline

The cascaded system described above uses SGMSE+ as a post-processing module. However, the iterative denoising process of diffusion-based models introduces additional computational cost during training and inference. This makes efficient end-to-end joint optimization with the separation network challenging.

To enable efficient joint optimization, we further explore an integrated training strategy. For this setting, the diffusion-based dereverberation model is replaced with SkipConvNet [33], a discriminative network with a stacked convolutional architecture. Its deterministic forward pass makes it more suitable for gradient-based joint training with the separation network. We then jointly train the separation module and the SkipConvNet-based dereverberation module. The total loss is defined as:

$$\mathcal{L}_{\text{joint}} = \mathcal{L}_{\text{separate}} + \mathcal{L}_{\text{dereverb}}$$

where $\mathcal{L}_{\text{separate}}$ is the progressive separation loss defined in Section 2.2, and $\mathcal{L}_{\text{dereverb}}$ is the mean squared error (MSE) loss on the magnitude spectrogram used to train SkipConvNet. Since the two losses are defined in different domains, this formulation should be interpreted as an empirical joint-training objective rather than a theoretically derived optimum. Training both modules jointly from scratch can lead to unstable convergence in our preliminary experiments. To improve training stability while still allowing adaptation between modules, we initialize the dereverberation module with pre-trained weights and then jointly train the full network.

3 Experimental setup

3.1 Datasets

To benchmark the AVSE system under realistic conditions, the 4th COG-MHEAR AVSEC [5] released a dataset featuring complex acoustic environments. Distinct from clean datasets such as LRS3 [34] and VoxCeleb2 [35], AVSEC-4 introduces several key complexities, including reverberation simulated using real room impulse responses, overlapping speech from two or more speakers, and non-stationary music noise as background interference.

The dataset is split into training (34,524 utterances, 113 hours), development (3,306 utterances, 8.6 hours), and test (3,180 utterances, 8.3 hours) sets. From

[5] `https://challenge.cogmhear.org/`

the test set, 1,500 utterances are designated for objective leaderboard evaluation, while the remaining 1,680 are reserved for subjective listening tests. The test set is simulated to reflect real conference room acoustics, with variability in room dimensions and source-receiver distances. The overall SNR spans from -18 dB to 6.55 dB, encompassing a wide range of challenging but realistic acoustic conditions.

3.2 Implementation Details

During training, we dynamically generate mixtures for each epoch. The target binaural speech is mixed with interfering speakers and noise with the SNR uniformly sampled between -18 dB and 6 dB. Both the generated mixture audio and the corresponding target-speaker video (at 25 fps) are then randomly cropped into 3-second segments.

The model is optimized using the Adam optimizer with an initial learning rate of 0.001. We employ a learning rate scheduler that halves the rate if the validation loss does not decrease for 3 consecutive epochs. Training is terminated early if no improvement is observed for 10 consecutive epochs to mitigate the risk of overfitting. The SI-SDR [27] is used as the training objective for the separation network. Due to memory constraints, training is conducted with a batch size of 2.

To further optimize the model, we explored an enhanced training strategy. Under this refined strategy, we extended the patience periods for the learning rate scheduler and early stopping to 15 and 30 epochs, respectively. This provides the model more time to converge stably. Furthermore, we introduce a multi-stage loss formulation. Once the learning rate decays to one-tenth of its initial value, we incorporate a supplementary STFT loss [36] to form a composite objective. The STFT loss imposes explicit spectral-domain constraints and is used to complement the SI-SDR objective. As discussed in Section 4, the complete-test-set analysis isolates the separation stage and compares two models that both take binaural mixtures as input but differ in their output-channel configuration. The subsequent ablation systems retain the same binaural input and produce monaural outputs, because the official clean reference available for the complete test set is monaural.

4 Results and Analysis

The objective results on the official 1,500-utterance leaderboard subset are summarized in Table 1. Because the submissions use different channel settings, the table groups monaural and binaural systems separately. The submitted system combines the same binaural separation model evaluated in the second row of Table 2 with SGMSE+ dereverberation. Among the reported binaural submissions, it obtains the highest PESQ, STOI, and SI-SDR scores, with values of 2.16, 0.82, and -17.47 dB, respectively. Table 1 reports the final separation–dereverberation output on a subset of the test data, whereas Tables 2 and 3 use the complete

Table 1: Official AVSEC-4 leaderboard results on the 1,500-utterance objective-evaluation subset. Results are grouped by channel setting, and bold values indicate the best score within each group. The submitted system (ours) uses the same binaural separation model evaluated in Table 2, followed by SGMSE+ dereverberation.

System	PESQ	STOI	SI-SDR
Monaural submissions			
Noisy mixture	1.29	0.51	-25.94
Team-OPTIMAL	1.37	0.48	-21.44
USTC_Entry1	1.35	0.52	-23.96
GU-ENU	1.35	0.52	-23.93
Rahma_Team	1.30	0.55	-24.72
TeamKCW	1.58	0.60	-21.77
SND_VD	1.72	0.64	-21.96
BioASP	1.71	0.67	-20.41
CITISIN	**2.29**	**0.78**	**-17.15**
Binaural submissions			
Noisy mixture	1.31	0.55	-24.28
R-test1	1.64	0.66	-19.26
SUSTechAILab	1.95	0.67	-18.75
ours	**2.16**	**0.82**	**-17.47**

Table 2: Separation-only comparison on the complete test set. Both separation models take the binaural mixture as input and differ only in their output-channel configuration; no dereverberation module is used. For the binaural mixture and binaural separation output, each channel is evaluated separately against the same official monaural target, and the two channel-wise scores are averaged. The monaural separation output is evaluated directly against the monaural target.

System	PESQ	STOI	SI-SDR
Binaural mixture (avg.)	1.08	0.51	-25.52
Binaural separation output (avg.)	1.42	0.78	-18.09
Monaural separation output	1.42	0.78	-18.22

test set and serve different analytical purposes. Their absolute metric values are therefore not directly comparable.

A formal subjective listening test was conducted by the challenge organizers. The subjective evaluation included both monaural and binaural listening tests, each using the same set of video clips. Each participant evaluated 126 clips, consisting of 120 test samples and 6 validation questions. The final word accuracy for each system was computed according to the official evaluation protocol. The results are depicted in Figure 3; the letters on the horizontal axis denote anonymized challenge entries, and our binaural submission is labeled "U". Our system achieved a word accuracy of 85.23% in the binaural subjective evaluation. This result indicates that the proposed system can improve perceived intelligibility under the AVSEC-4 subjective evaluation setting. However, since statistical confidence intervals are not provided in the official figure, we avoid making claims about statistical significance among systems.

Table 2 isolates the separation stage on the complete test set and is intended to justify the output-channel setting used in the subsequent ablation study. Both separation models take the same binaural mixture as input. The binaural-output model is identical to the separation stage of the submitted system in Table 1, but Table 2 excludes SGMSE+ dereverberation. Because the official clean reference for the complete test set is monaural, each channel of the binaural mixture and binaural separation output is evaluated independently against the same monaural target, after which the two scores are averaged. The monaural-output separation model is evaluated directly against the monaural target. Under this protocol, the two separation models obtain the same PESQ and STOI values at the reported precision, while their SI-SDR scores differ by 0.13 dB. These results show that the two output configurations yield similar objective scores under the adopted monaural-reference protocol; they do not establish general equivalence between binaural and monaural processing or evaluate binaural spatial fidelity. We therefore use the binaural-input, monaural-output separation model in the following ablation experiments so that all system outputs can be evaluated directly against the official monaural clean reference. It is noticable that, since Table 1 uses a different test subset and includes SGMSE+ post-processing, its absolute scores should not be compared directly with those in Table 2.

Table 3 reports the ablation study on the complete test set. All ablation systems retain the binaural mixture as input and use the monaural-output configuration validated in Table 2; their outputs are therefore evaluated directly against the official monaural clean reference. The "Separation only" row is the same monaural separation output reported in Table 2. The binaural mixture is included only as the unprocessed input reference, and its score is obtained by evaluating the two channels independently against the same monaural target and averaging the resulting values.

Among the non-joint SGMSE+-based configurations, applying dereverberation before separation yields lower scores than the separation-based systems. This result is consistent with the possibility that processing the mixture before target-speaker extraction alters information useful for the subsequent separation

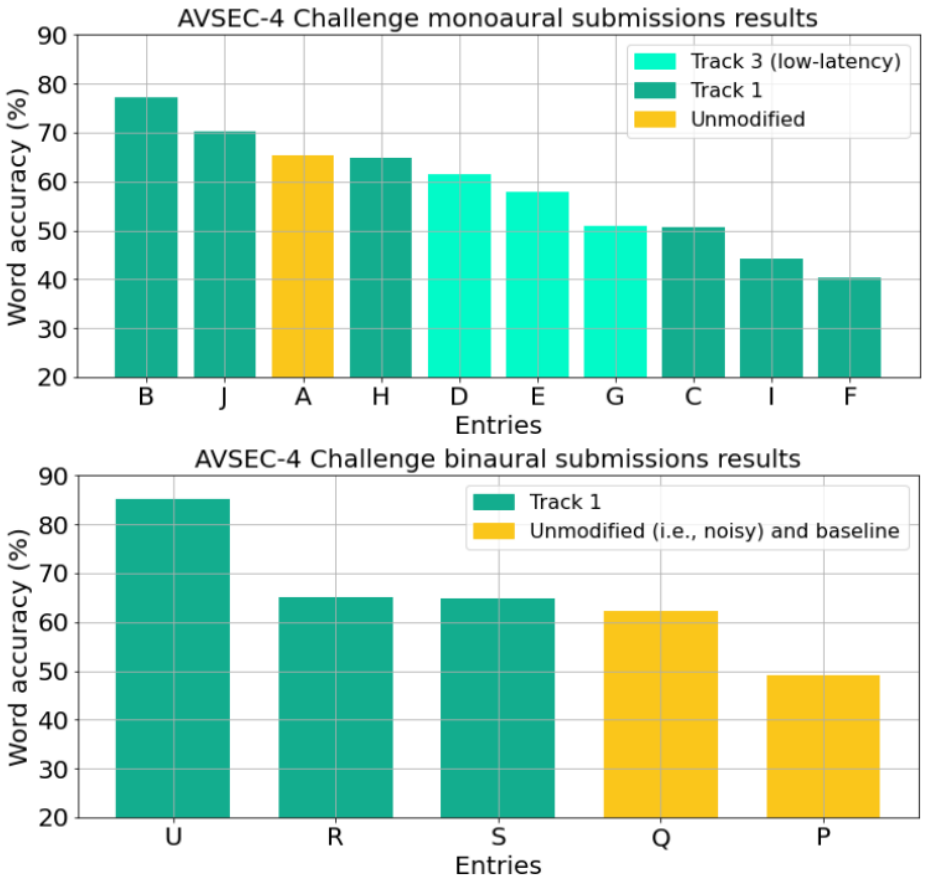


Fig. 3: Subjective evaluation results reported by the AVSEC-4 challenge organizers. The x-axis labels denote anonymized challenge entries. Our binaural submission is represented by "U".

Table 3: Ablation results on the complete test set. All ablation systems take the binaural mixture as input, produce monaural outputs, and are evaluated directly against the official monaural clean reference. The binaural mixture row is an unprocessed input reference obtained by averaging the two channel-wise scores, as in Table 2. Non-joint cascaded systems use SGMSE+, whereas the jointly trained system uses SkipConvNet.

System	PESQ	STOI	SI-SDR
Binaural mixture (avg.)	1.08	0.51	-25.52
SGMSE+ before separation	1.35	0.69	-23.71
Separation only	1.42	0.78	-18.22
Separation before SGMSE+	1.44	0.81	-18.71
Separation only (joint model)	1.44	0.79	-18.07
Separation before SkipConvNet (joint)	**1.69**	**0.83**	**-17.66**

stage, although the present experiments do not directly establish the underlying mechanism. Applying SGMSE+ after separation improves PESQ and STOI over the separation-only model, while its SI-SDR is slightly lower. Thus, under the tested configuration, the post-processing stage benefits the perceptual-quality and intelligibility-related metrics but does not improve the waveform-level distortion metric.

The jointly trained SkipConvNet-based configuration obtains the highest values in Table 3. The "Separation only (joint model)" row evaluates the intermediate separation output of this jointly trained system before SkipConvNet processing. However, the jointly trained system differs from the cascaded SGMSE+-based system in both the dereverberation architecture and the optimization strategy. Consequently, the observed gain cannot be attributed exclusively to joint training. Moreover, a jointly trained "dereverberation before separation" configuration is not available. The conclusions are therefore restricted to the reported systems rather than a complete factorial comparison of module order, dereverberation architecture, and training strategy.

5 Conclusion

In conclusion, we present a system-level AVSE framework that combines binaural audio-visual separation, progressive loss training, and dereverberation for complex acoustic environments. The submitted AVSEC-4 system uses a binaural separation model followed by SGMSE+ and achieves competitive objective results together with a high word-accuracy score in the subjective listening evaluation. On the complete test set, the binaural-output and monaural-output separation models yield similar objective scores under the adopted monaural-reference protocol, which supports the use of monaural outputs for the subsequent controlled analysis while retaining binaural mixtures as input. Among the tested non-joint configurations, applying SGMSE+ after separation improves PESQ and STOI relative to separation alone. The jointly trained SkipConvNet-based system obtains the highest reported values, although its difference from the SGMSE+-based cascade cannot be attributed solely to joint optimization because the dereverberation architecture also changes. The current study is therefore primarily a system-level integration and validation; additional visual-cue ablations, complexity measurements, and matched dereverberation comparisons would be required to establish broader conclusions.

Acknowledgements This research is funded in part by the National Natural Science Foundation of China (62571223) and Yangtze River Delta Science and Technology Innovation Community Joint Research Project (2024CSJGG01100).

References

1. Wang, J., Pan, Z., Zhang, M., Tan, R.T., Li, H.: Restoring speaking lips from occlusion for audio-visual speech recognition. In: Proceedings of the AAAI conference on artificial intelligence, vol. 38, pp. 19144–19152 (2024)

2. Arons, B.: A review of the cocktail party effect. Journal of the American Voice I/O society **12**(7), 35–50 (1992)
3. Hershey, J.R., Chen, Z., Le Roux, J., Watanabe, S.: Deep clustering: Discriminative embeddings for segmentation and separation. In: Proc.ICASSP, pp. 31–35 (2016)
4. Luo, Y., Mesgarani, N.: Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. IEEE/ACM transactions on audio, speech, and language processing **27**(8), 1256–1266 (2019)
5. Wang, Z.Q., Cornell, S., Choi, S., Lee, Y., Kim, B.Y., Watanabe, S.: TF-GridNet: Integrating full-and sub-band modeling for speech separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing **31**, 3221–3236 (2023)
6. Muckenhirn, H.R., Moreno, I.L., Hershey, J., Wilson, K., Sridhar, P., Wang, Q., Saurous, R.A., Weiss, R., Jia, Y., Wu, Z.: Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking. In: Proc.Interspeech (2019)
7. Xu, C., Rao, W., Chng, E.S., Li, H.: Spex: Multi-scale time domain speaker extraction network. IEEE/ACM transactions on audio, speech, and language processing **28**, 1370–1384 (2020)
8. Ge, M., Xu, C., Wang, L., Chng, E.S., Dang, J., Li, H.: Spex+: A complete time domain speaker extraction network. In: Proc. Interspeech, pp. 1406–1410 (2020)
9. Hao, F., Li, X., Zheng, C.: X-TF-GridNet: A time–frequency domain target speaker extraction network with adaptive speaker embedding fusion. Information Fusion **112**, 102550 (2024)
10. Michelsanti, D., Tan, Z.H., Zhang, S.X., Xu, Y., Yu, M., Yu, D., Jensen, J.: An overview of deep-learning-based audio-visual speech enhancement and separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing **29**, 1368–1396 (2021)
11. Gu, R., Zhang, S.X., Xu, Y., Chen, L., Zou, Y., Yu, D.: Multi-modal multi-channel target speech separation. IEEE Journal of Selected Topics in Signal Processing **14**(3), 530–541 (2020)
12. Ceolini, E., Hjortkjær, J., Wong, D.D., O'Sullivan, J., Raghavan, V.S., Herrero, J., Mehta, A.D., Liu, S.C., Mesgarani, N.: Brain-informed speech separation (BISS) for enhancement of target speaker in multitalker speech perception. NeuroImage **223**, 117282 (2020)
13. Pan, Z., Wang, W., Borsdorf, M., Li, H.: ImagineNet: Target speaker extraction with intermittent visual cue through embedding inpainting. In: Proc.ICASSP, pp. 1–5 (2023)
14. Pan, Z., Borsdorf, M., Cai, S., Schultz, T., Li, H.: NeuroHeed: Neuro-steered speaker extraction using EEG signals. IEEE/ACM Transactions on Audio, Speech, and Language Processing (2024)
15. Martinez, B., Ma, P., Petridis, S., Pantic, M.: Lipreading using temporal convolutional networks. In: Proc. ICASSP, pp. 6319–6323 (2020)
16. Fenghour, S., Chen, D., Guo, K., Li, B., Xiao, P.: Deep learning-based automated lip-reading: A survey. IEEE Access **9**, 121184–121205 (2021)
17. Wu, J., Xu, Y., Zhang, S.X., Chen, L.W., Yu, M., Xie, L., Yu, D.: Time domain audio visual speech separation. In: Proc. ASRU, pp. 667–673 (2019)
18. Pan, Z., Tao, R., Xu, C., Li, H.: Muse: Multi-modal target speaker extraction with visual cues. In: Proc. ICASSP, pp. 6678–6682 (2021)
19. Pan, Z., Ge, M., Li, H.: USEV: Universal speaker extraction with visual cue. IEEE/ACM Transactions on Audio, Speech, and Language Processing **30**, 3032–3045 (2022)
20. Pan, Z., Qian, X., Li, H.: Speaker extraction with co-speech gestures cue. IEEE Signal Processing Letters **29**, 1467–1471 (2022)

21. Li, K., Xie, F., Chen, H., Yuan, K., Hu, X.: An audio-visual speech separation model inspired by cortico-thalamo-cortical circuits. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(10), 6637–6651 (2024)
22. Li, K., Yang, R., Sun, F., Hu, X.: IIANet: an intra-and inter-modality attention network for audio-visual speech separation. In: Proc. ICML, pp. 29181–29200 (2024)
23. Jin, Z., Zeng, B., Li, Z., Liu, X., Li, M.: A Target Speaker Extraction Method for the 3rd Audio-Visual Speech Enhancement Challenge. In: Proc. AVSEC 2024, pp. 16–18 (2024)
24. Hou, Z., Sun, T., Hu, Y., Zhu, C., Chen, K., Lu, J.: Sir-progressive audio-visual tf-gridnet with ASR-aware selector for target speaker extraction in MISP 2023 challenge. In: Proc. ICASSPW, pp. 11–12 (2024)
25. Rix, A.W., Beerends, J.G., Hollier, M.P., Hekstra, A.P.: Perceptual evaluation of speech quality (PESQ): A new method for speech quality assessment of telephone networks and codecs. In: 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221), vol. 2, pp. 749–752. IEEE (2001)
26. Taal, C.H., Hendriks, R.C., Heusdens, R., Jensen, J.: A short-time objective intelligibility measure for time-frequency weighted noisy speech. In: Proc. ICASSP, pp. 4214–4217 (2010)
27. Le Roux, J., Wisdom, S., Erdogan, H., Hershey, J.R.: SDR–half-baked or well done?. In: Proc. ICASSP, pp. 626–630 (2019)
28. Pan, Z., Wichern, G., Masuyama, Y., Germain, F.G., Khurana, S., Hori, C., Le Roux, J.: Scenario-aware audio-visual TF-Gridnet for target speech extraction. In: Proc. ASRU, pp. 1–8 (2023)
29. Jin, Z., Zeng, B., Yang, P., Du, J., Liu, J., Li, M.: Robust audio-visual target speaker extraction with emotion-aware multiple enrollment fusion. arxiv 2509.12583 (2025)
30. Roy, A.K., Kathania, H.K., Sharma, A., Dey, A., Ansari, M.S.A.: ResEmoteNet: Bridging Accuracy and Loss Reduction in Facial Emotion Recognition. IEEE Signal Processing Letters **32**, 491-495 (2024)
31. Zhao, Z., Yang, D., Gu, R., Zhang, H., Zou, Y.: Target Confusion in End-to-end Speaker Extraction: Analysis and Approaches. In: Proc. Interspeech, pp. 5333–5337 (2022)
32. Richter, J., Welker, S., Lemercier, J.M., Lay, B., Gerkmann, T.: Speech enhancement and dereverberation with diffusion-based generative models. IEEE/ACM Transactions on Audio, Speech, and Language Processing **31**, 2351–2364 (2023)
33. Kothapally, V., Xia, W., Ghorbani, S., Hansen, J.H., Xue, W., Huang, J.: SkipConvNet: Skip Convolutional Neural Network for Speech Dereverberation Using Optimally Smoothed Spectral Mapping. In: Proc. Interspeech, pp. 3935–3939 (2020)
34. Afouras, T., Chung, J.S., Zisserman, A.: LRS3-TED: a large-scale dataset for visual speech recognition. arXiv:1809.00496 (2018)
35. Chung, J., Nagrani, A., Zisserman, A.: VoxCeleb2: Deep speaker recognition. in Proc. Interspeech 1086-1090 (2018)
36. Pan, Z., Ge, M., Li, H.: A Hybrid Continuity Loss to Reduce Over-Suppression for Time-domain Target Speaker Extraction. In: Proc. Interspeech, pp. 1786–1790 (2022)
37. Wang, Q., Muckenhirn, H., Wilson, K., Sridhar, P., Wu, Z., Hershey, J., Saurous, R.A., Weiss, R.J., Jia, Y., Moreno, I.L.: Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking. arXiv preprint arXiv:1810.04826 (2018)
38. Zeng, B., Suo, H., Wan, Y., Li, M.: Sef-net: Speaker embedding free target speaker extraction network. In: Proc. Interspeech, pp. 3452–3456 (2023)

39. Li, A., Zheng, C., Peng, R., Li, X.: On the importance of power compression and phase estimation in monaural speech dereverberation. JASA express letters **1**(1) (2021)
40. Tu, Y.H., Du, J., Gao, T., Lee, C.H.: A multi-target SNR-progressive learning approach to regression based speech enhancement. IEEE/ACM Transactions on Audio, Speech, and Language Processing **28**, 1608–1619 (2020)
41. Valin, J.M., Giri, R., Venkataramani, S., Isik, U., Krishnaswamy, A.: To dereverb or not to dereverb? Perceptual studies on real-time dereverberation targets. arXiv preprint arXiv:2206.07917 (2022)
42. Tao, R., Qian, X., Jiang, Y., Li, J., Wang, J., Li, H.: Audio-visual target speaker extraction with selective auditory attention. IEEE Transactions on Audio, Speech and Language Processing (2025)
43. Shi, B., Hsu, W.N., Lakhotia, K., Mohamed, A.: Learning audio-visual speech representation by masked multimodal cluster prediction. arXiv preprint arXiv:2201.02184 (2022)
44. Wu, W., Chen, X., Wu, X., Li, H., Meng, H.: Target speech extraction with pre-trained av-hubert and mask-and-recover strategy. In: 2024 International Joint Conference on Neural Networks (IJCNN), pp. 1–8 (2024)
45. Shi, B., Hsu, W.N., Mohamed, A.: Robust self-supervised audio-visual speech recognition. arXiv preprint arXiv:2201.01763 (2022)
46. Shi, B., Mohamed, A., Hsu, W.N.: Learning lip-based audio-visual speaker embeddings with av-hubert. arXiv preprint arXiv:2205.07180 (2022)
47. Chern, I.C., Hung, K.H., Chen, Y.T., Hussain, T., Gogate, M., Hussain, A., Tsao, Y., Hou, J.C.: Audio-visual speech enhancement and separation by utilizing multi-modal self-supervised embeddings. In: 2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW), pp. 1–5 (2023)
48. Assael, Y.M., Shillingford, B., Whiteson, S., De Freitas, N.: Lipnet: End-to-end sentence-level lipreading. arXiv preprint arXiv:1611.01599 (2016)
49. Fung, I., Mak, B.: End-to-end low-resource lip-reading with maxout CNN and LSTM. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2511–2515 (2018)
50. Torfi, A., Iranmanesh, S.M., Nasrabadi, N., Dawson, J.: 3D convolutional neural networks for cross audio-visual matching recognition. IEEE Access **5**, 22081–22091 (2017)
51. Son Chung, J., Senior, A., Vinyals, O., Zisserman, A.: Lip reading sentences in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6447–6456 (2017)
52. Stafylakis, T., Tzimiropoulos, G.: Combining residual networks with LSTMs for lipreading. arXiv preprint arXiv:1703.04105 (2017)
53. Petridis, S., Stafylakis, T., Ma, P., Cai, F., Tzimiropoulos, G., Pantic, M.: End-to-end audiovisual speech recognition. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 6548–6552 (2018)
54. Margam, D.K., Aralikatti, R., Sharma, T., Thanda, A., Roy, S., Venkatesan, S.M., others: LipReading with 3D-2D-CNN BLSTM-HMM and word-CTC models. arXiv preprint arXiv:1906.12170 (2019)
55. Zeng, B., Li, M.: Usef-tse: Universal speaker embedding free target speaker extraction. IEEE Transactions on Audio, Speech and Language Processing (2025)
56. Zeghidour, N., Grangier, D.: Wavesplit: End-to-end speech separation by speaker clustering. IEEE/ACM Transactions on Audio, Speech, and Language Processing **29**, 2840–2849 (2021)