Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

Online AV-CrossNet: a Causal and Efficient Audiovisual System for Speech Enhancement and Target Speaker Extraction

Cheng Yu[1], *Vahid Ahmadi Kalkhorani*[1], *Buye Xu*[3], *and DeLiang Wang*[1,2]

[1]Department of Computer Science and Engineering, The Ohio State University, USA
[2]Center for Cognitive and Brain Sciences, The Ohio State University, USA
[3]Meta Reality Labs, USA

{yu.3500, ahmadikalkhorani.1, wang.77}@osu.edu , xub@meta.com

Abstract

This paper presents online AV-CrossNet, a computationally efficient audiovisual speech enhancement/extraction system capable of causal and real-time processing. We aim to improve the state-of-the-art AV-CrossNet by enabling causal, frame-by-frame processing. To achieve this, we incorporate causal layers and compression techniques, reduce model size, and employ only one-frame look-ahead, thereby substantially enhancing real-world applicability. Additionally, we analyze compression ratio in both audio and visual modules, providing valuable insights into audiovisual model compression. Experimental results demonstrate an inference latency of 4.73 ms, capable of real-time processing. Moreover, the system maintains competitive performance while reducing size by a factor of 10. These findings highlight the efficiency and effectiveness of the proposed system, offering a promising solution for real-time audiovisual speech enhancement and speaker extraction in acoustically adverse environments.

Index Terms: Audiovisual speech enhancement, audiovisual target speaker extraction, TF-CrossNet, AV-CrossNet

1. Introduction

In real-world environments, speech signals are corrupted by various types of noise, leading to significant degradation in both speech quality and intelligibility. Recently, the deep learning based speech separation systems have demonstrated remarkable performance across a wide range of scenarios [1]. Despite these advances, existing systems face limitations in extremely noisy conditions [2] or when the noisy mixture includes speech interference, such as in meetings and conversations [3].

Recent advancements in audiovisual (AV) speech processing have led to significant improvements in both audiovisual speech enhancement (AVSE) and audiovisual target speaker extraction (AVTSE) tasks [4]. These approaches leverage both visual and auditory cues, with the visual information associated with talking unaffected by acoustic interference [5]. By integrating multimodal features, these AV separation methods effectively overcome the limitations of audio-only models, particularly in adverse acoustic environments that severely degrade acoustic cues [6–12]. Despite these advance, a notable gap remains in the application of these techniques for real-world deployment. A main reason is that these high-performing AV algorithms are not designed to support online inference, a crucial requirement for real-time applications.

A real-time system must maintain algorithmic causality and computational efficiency. While some studeis have addressed real-time AVSE on edge devices [13, 14], these studies do not evaluate real-time performance. Others have demonstrated their systems on the server side. Zhu et al. [15] proposed an LSTM-based AVSE system with multi-stage fusion and an efficient visual encoder. Montesinos et al. [16] introduced a transformer-based AV voice separation system. Gu et al. [17] presented a multi-channel audiovisual speaker separation (AVSS) system. However, these approaches do not compare between causal and non-causal systems, and hence the performance cost of causal implementation is unclear. Chen et al. [18] proposed a resynthesis based AVSE and conducted extensive evaluations of system efficiency and effectiveness, including comparisons to non-causal systems. While these methods have demonstrated real-time processing, they do not address both AVSE and AVTSE. Furthermore, deep model compression is not investigated, and a small model is often necessary for ensuring real-world applicability across various devices.

To address these challenges, we propose online AVCrossNet, which builds on the AV-CrossNet model [19] by introducing causal processing. Specifically, online AV-CrossNet is a causal and efficient AV system that can performs both AVSE and AVTSE tasks on the basis of frame-by-frame complex spectral mapping. We also perform model compression on the proposed system, exploring the compression of auditory and visual modules. Experimental results show that our proposed system maintains high performance compared to the non-causal counterparts in various conditions. In addition, we are able to compress our system up to 10 times while retaining competitive AVSE and AVTSE results.

2. Methodology

This section describes our system, including its architecture and loss functions. Our system is extended from the AV-CrossNet model [19]. We introduce several key improvements to the original architecture to enable online inference capabilities.

2.1. Algorithmic Causality

AV-CrossNet is algorithmically non-causal. Both the visual encoder [20] and TF-CrossNet blocks [21] are designed to leverage future information to achieve optimal performance. Specifically, the convolutional layers require up to three frames of look-ahead, while the global multi-head self-attention (GMHSA) module depends on both past and future features within a data batch. Consequently, directly applying online inference to non-causal systems would lead to a significant performance degradation [18]. We propose online AV-CrossNet with algorithmic 1-frame look-ahead (40 ms). Figure 1 shows a diagram of the proposed online system, and its details are described as follows.

10.21437/Interspeech.2025-1448

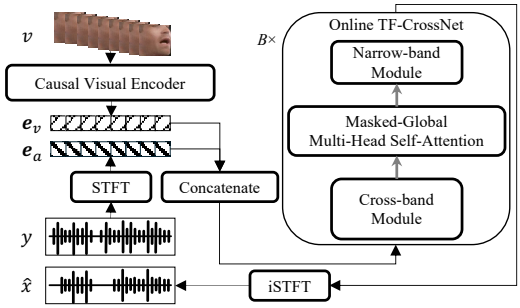

Figure 1: *Diagram of the proposed online AV-CrossNet. The auditory and visual input, $\mathbf{y}$ and $\mathbf{v}$ are encoded by short-time Fourier transform (STFT) and a causal visual encoder respecitvely, generating the embeddings $\mathbf{e}_a$ and $\mathbf{e}_v$. Then, the online TF-CrossNet blocks takes the concatenated embeddings and predict for the enhanced speech $\hat{x}$. B denotes the number of blocks.*

2.1.1. Causal Visual Encoder

As shown in Figure 1, online AV-CrossNet includes a causal visual encoder. This encoder consists of three modules: a 3-D causal convolutional encoder, a causal ResNet-18 model, and a five-layered causal visual temporal convolutional network (V-TCN) block to capture long-term feature dependencies [22]. In the original visual encoder, the convolutional layers require a look-ahead of up to three video frames, which violates algorithmic causality. Additionally, the batch normalization layers depend on the statistics from all time frames during training. To make this encoder algorithmically causal, we introduce two modifications.

First, we replace each convolutional layer with a causal convolutional layer. These layers apply left-padding to the features along the time axis according to the kernel size [23]. Specifically, for an arbitrary convolutional layer with kernel size k along the time axis, the padding size is $\lfloor \frac{k}{2} \rfloor$, where $\lfloor \cdot \rfloor$ denotes the floor operation. Second, we replace batch normalization layers with group normalization layers [24]. The group normalization layer is independent of batch size and maintains algorithmic causality, providing performance comparable to that of batch normalization. Unlike AV-CrossNet, which relies on a pretrained visual encoder from DeepAVSR [20], our system cannot utilize pretrained encoder due to its causal model design, which does not permit pretraining in the same way as non-causal models. In addition to model architecture, we introduce causal linear interpolation to maintain algorithmic causality in the video stream. Specifically, for the first visual cue, we replicate it until the system receives the subsequent frame. Then, we perform linear interpolation between the past and present frame to ensure a smooth transition of visual cues along the temporal axis, until the last frame.

2.1.2. Online TF-CrossNet

As illustrated in Figure 1, the online TF-CrossNet block consists of three modules: a masked-GMHSA module, a cross-band module, and a narrow-band module. To strike a balance between system efficiency and performance, we chose not to include the narrow-band multi-head self-attention (MHA) module [19]. We design the masked-GMHSA module by implementing a causal self-attention mechanism, which applies a lower triangular mask to the attention logits, thereby preventing future information from influencing the attention computation. For the rest of the architecture, we replace convolutional layers with causal convolutional layers. To enhance efficiency and effectiveness, we introduce a variant of the narrow-band module that incorporates the Mamba state-space model [25], which has gained prominence in speech processing. The Mamba layer, built from causal convolutional layers, is effective and efficient in processing complex spectral features [26]. The Mamba narrow-band module enhances both training and inference efficiency, slightly decreases the model size, and improves overall performance. To differentiate this version from the original version, we refer to it as online TF-CrossNet-Mamba.

2.2. Model Compression

To enhance the online applicability of our system, we further investigate model compression techniques, including weight pruning and quantization [27]. Weight pruning identifies and removes weights with lower values within a weight group, assuming these weights contribute less to the model's performance. Weight quantization clusters weights and replaces them within each cluster with their core value, which is then stored in a codebook. In this work, we apply unstructured weight pruning and quantization to our proposed system. We define weight groups in the proposed systems following [28]. The weights are stored as 32-bit floating-point numbers. The overall compression rate r for a system with N total weights, N_p non-zero weights after pruning and clustered into K clusters, can be computed as follows:

$$r = \frac{32N}{N_p \log_2 K + 32K} \tag{1}$$

In this study, we adopt the 16-bit precision for inference, which significantly improves the system's efficiency.

2.3. Loss Functions

We use the hybrid loss function [29] to train all models. This hybrid loss combines the scale-invariant signal-to-distortion ratio (SI-SDR) loss, which optimizes signal-level performance, and the multi-resolution STFT loss, which ensures accurate spectral mapping.

3. Experimental Setup

This section provides details of the preparation of the proposed system, including datasets and hyper-parameters.

3.1. Datasets

In this work, we propose online AV-CrossNet to tackle two related tasks: AVSE and AVTSE. For each task, we employ distinct datasets for training, validation, and test. The video and audio modalities in these datasets are synchronized, with video frame rate of 25 frames per second and audio sampling rate of 16,000 Hz.

3.1.1. Audiovisual Speech Enhancement

For AVSE, we prepare LRS3-AudioSet and COG-MHEAR AVSE Challenge [12] datasets. The LRS3-AudioSet is composed of speech signals from LRS3 dataset [30], which consists of about 100k, 10k, and 2.1k utterances for training, validation, and test. The noise signals are from the non-speech subset of the AudioSet [31], which consists of 2.7k, 305, and 86 signals for training, validation, and test. For training and validation, we adopt the dynamic mixing with signal-to-noise ratio (SNR)

range of $[-15,\ 15]$ dB, and truncate all utterances into 3 seconds. For test set, we prepare 2.1k utterances in 7 SNR levels in the range of $[-15,\ 15]$ dB, with 5 dB gap between each level.

The COG-MHEAR AVSE Challenge dataset uses speech signals from the LRS3 dataset, and noise signals from the Clarity challenge [32], Demand [33], and Freesound datasets [34]. Following [19], we reorganize this dataset into training, validation, and test sets. The validation set includes 1,339 utterances for AVSE and 1,358 for AVTSE, selected from the 24 most frequent speakers in the training set. The training, validation, and test sets contain 30,297, 3,017, and 3,306 utterances, respectively. Training and validation utterances are truncated to 3 seconds, while test utterances vary from 2 to 30 seconds. This dataset adopts speech-weighted SNRs sampled in the range of $[-10,\ 10]$ dB for AVSE and $[-15,\ 5]$ dB for AVTSE.

3.1.2. Audiovisual Target Speaker Extraction

Following [7–10, 19], we evaluate the proposed system on the AVTSE task under various conditions. All systems are trained on the VoxCeleb2 dataset [35], with 48,000 utterances from 800 speakers for training, and 36,237 utterances from 118 speakers for test, ensuring no overlap between training, validation, and test speakers. All utterances are at least 4 seconds long. The training, validation, and test sets contain 20,000, 5,000, and 3,000 mixtures of two speakers, respectively. For the mismatched evaluation, we use the LRS3 and TCD-TIMIT [36] datasets, generating 3,000 mixtures by randomly mixing target speech with a competing speaker. The target-to-interferer ratio (TIR) is randomly sampled in the range of $[-10,\ 10]$ dB.

3.2. Hyperparameters

For the AVSE task, we prepare a system consisting of twelve online TF-CrossNet blocks (B = 12). For the AVTSE task, we prepare a smaller system consisting of six online TF-CrossNet blocks (B = 6). The rest of the modules remain the same in both systems. For each training, we utilize Adam optimizer with a maximum learning rate of 0.001. We apply the ReduceLROnPlateau learning rate scheduler with a decay factor of 0.85 and patience of 3 epochs. We adopt the SNR-scheduler [37] for the COG-MHEAR AVSE Challenge dataset, starting with SNR range of $[-15,\ -5]$ dB for AVSE and $[-20,\ -10]$ dB for AVTSE. The scheduler increases both the upper and lower bounds by 5 dB if no validation improvement is observed for 6 consecutive epochs. The scheduler stops when it reaches the objective SNR range, which is $[-10,\ 10]$ dB for AVSE and $[-15,\ 5]$ dB for AVTSE. Each system is trained with a batch size that maximizes GPU memory usage. Training continues until no further improvement in validation loss is observed over 20 epochs.

4. Experimental Results

In this section, we present the experimental results for the proposed online TF-CrossNet (oTF-CrossNet) model, online AV-CrossNet (oAV-CrossNet) model and online AV-CrossNet-Mamba (oAV-CrossNet-Mamba) model in both AVSE and AVTSE scenarios. Additionally, we provide a model compression analysis of the proposed system.

4.1. Audiovisual Speech Enhancement

In Table 1, we evaluate and compare our proposed systems, oTF-CrossNet, oTF-CrossNet-Mamba, oAV-CrossNet, and oAV-CrossNet-Mamba, with several baseline models on AVSE task. The audio-only oTF-CrossNet outperforms both audio-only and AV HD-Demucs [38]. However, it falls short of oSpatialNet-Mamba [26]. In contrast, the proposed oTF-CrossNet-Mamba slightly outperforms oSpatialNet-Mamba in both perceptual evaluation of speech quality (PESQ) and SI-SDR metrics. This advantage is from GMHSA module, which models global time-frequency feature more effectively [21]. Furthermore, both oTF-CrossNet-Mamba and oAV-CrossNet-Mamba outperform their generic counterparts, confirming the Mamba layer's effectiveness in processing time-frequency features. Moreover, the performance degradation from the non-causal systems to the proposed systems is within a reasonable range for the AVSE task, as reported in [18].

Table 1: *Evaluation results of the proposed online AV-CrossNet and baseline models on the LRS3-Audio Set test set.*

Method	AV	PESQ	STOI	SI-SDR
Unprocessed	-	1.32	0.72	0.08
Offline Systems				
TF-CrossNet	✗	2.77	0.90	13.90
AV-CrossNet	✓	**2.82**	**0.92**	**13.61**
Online Systems				
HD-Demucs	✗	2.25	0.83	10.70
AV-HD-Demucs	✓	2.35	0.87	11.80
oSpatialnet-Mamba	✗	2.51	0.87	11.36
oTF-CrossNet	✗	2.45	0.86	11.22
oTF-CrossNet-Mamba	✗	2.53	0.87	11.49
oAV-CrossNet	✓	2.50	0.86	11.73
oAV-CrossNet-Mamba	✓	**2.60**	**0.90**	**12.32**

Table 2: *Evaluation results of the proposed online AV-CrossNet and baseline models on the development set for the second COG-MHEAR Audiovisual Speech Enhancement Challenge.*

Method	Speech+Noise			Speech+Speech		
	PESQ	STOI	SI-SDR	PESQ	STOI	SI-SDR
Unprocessed	1.15	0.68	-4.4	1.17	0.60	-5.0
Offline Systems						
AV-DPRNN	2.02	0.86	11.4	2.23	0.90	12.6
AV-GridNet	2.62	0.91	13.9	3.10	**0.95**	16.7
SAV-GridNet	2.68	0.91	14.2	**3.23**	**0.95**	**17.5**
AV-CrossNet	**2.75**	**0.92**	**14.3**	**3.23**	**0.95**	17.3
Online Systems						
oAV-CrossNet	2.35	0.88	12.7	2.61	0.92	14.3
oAV-CrossNet-Mamba	**2.41**	**0.89**	**13.1**	**2.72**	**0.93**	**14.9**

In Table 2, we evaluate our proposed system and compare with non-causal baseline models in both the AVSE and AVTSE tasks. The oAV-CrossNet-Mamba outperforms both oAV-CrossNet and the non-causal AV-DPRNN [39] in both tasks, demonstrating the superior performance of the Mamba narrowband module over the original module. However, it falls short of the more complex AV-GridNet and the ensemble system SAV-GridNet [11]. Furthermore, the performance degradation observed when transitioning from the non-causal to causal model is more pronounced in AVTSE task, indicating that AVTSE exhibits greater sensitivity to algorithmic causality than AVSE.

4.2. Audiovisual Target Speaker Extraction

We then evaluate our system on AVTSE task utilizing a smaller model configuration (B = 6). Table 3 presents a comparison of the systems under matched training and test conditions using the VoxCeleb2 dataset. Although the proposed systems underperform to AV-CrossNet, they outperform other non-causal baseline systems in terms of PESQ score, while requiring significantly fewer parameters, thereby demonstrating the effectiveness of the architecture. In Table 4, we evaluate the generalizability of the system. To investigate the role of the visual encoder, we prepare *AV-CrossNet-small in addition to the original AV-CrossNet, with the visual encoder trained with random initialization. The proposed online systems demonstrate competitive performance on the LRS3 dataset, but notably underperform on the TCD-TIMIT dataset. A similar trend is observed with *AV-CrossNet-small. These results suggest that the pre-trained visual encoder plays a crucial role in the system's generalizability on AVTSE task.

Table 3: *Evaluation results on VoxCeleb2 test set. Param is indicative of the overall system parameters, measured in millions (M). Each system is trained on the VoxCeleb2.*

Method	Param (M)	PESQ	SI-SDR
Unprocessed	-	-0.08	1.24
Offline Systems			
VisualVoice [8]	63.9	1.97	9.73
AV-ConvTasNet [7]	22.2	1.97	10.38
MuSE [9]	26.2	2.20	11.24
AV-SepFormer [10]	40.8	2.31	12.13
AV-CrossNet-small [19]	16.9	**2.93**	**14.71**
Online Systems			
oAV-CrossNet-small	16.9	2.31	**10.17**
oAV-CrossNet-Mamba-small	**16.5**	**2.34**	9.86

Table 4: *Evaluation results on mismatched test sets. Each system is trained on the VoxCeleb2.*

	LRS3		TCD-TIMIT	
Method	PESQ	SI-SDR	PESQ	SI-SDR
Unprocessed	1.21	0.13	1.47	-0.15
Offline Systems				
VisualVoice [8]	2.27	11.60	2.25	10.88
AV-ConvTasNet [7]	2.33	12.13	2.21	11.53
MuSE [9]	2.56	12.97	2.45	12.50
AV-SepFormer [10]	2.67	13.81	2.57	13.44
AV-CrossNet-small [19]	**3.14**	**17.42**	**3.25**	**18.15**
*AV-CrossNet-small	3.13	16.89	2.94	12.35
Online Systems				
oAV-CrossNet-small	2.50	12.75	2.13	4.34
oAV-CrossNet-Mamba-small	**2.54**	**13.13**	**2.17**	**4.82**

4.3. Model Compression

We examine model compression rates using oAV-CrossNet-Mamba on the COG-MHEAR dataset. In Tables 5 and 6, we report evaluation scores and the corresponding compression rates (r) with respect to the number of clusters for weight quantization (K) and the pruning ratios for the visual and audio modules (p_v, p_a), where a higher ratio means more pruning. The results indicate that, for both AVSE and AVTSE tasks, the visual module demonstrates higher sparsity compared to the audio module. Notably, the compressed model with K = 128, p_a = 0.1, and p_v = 0.9 outperforms the other compressed models that have p_a values between 0.2 and 0.3 and p_v values ranging from 0.5 to 0.7 on both PESQ and short-time objective intelligibility (STOI) metrics, while also achieving a substantially higher compression rates. Moreover, the AVTSE task experiences a more pronounced performance degradation than the AVSE task, suggesting its greater sensitivity to weight compression. Additionally, the AVTSE task exhibits more significant degradation as p_v increases, highlighting the pivotal role of the visual modality in this task.

Table 5: *Compression rates on the AVSE task*

	K=128			K=64		
p_a/p_v	PESQ	STOI	r	PESQ	STOI	r
0.0/0.0	**2.32**	**0.88**	4.6	**2.17**	**0.88**	5.3
0.1/0.5	2.14	**0.88**	6.2	2.08	0.87	7.2
0.1/0.7	2.25	**0.88**	7.4	1.84	0.85	8.6
0.1/0.9	2.21	**0.88**	9.2	1.72	0.83	10.8
0.2/0.5	2.16	**0.88**	6.5	1.84	0.86	7.6
0.2/0.7	2.16	**0.88**	7.9	1.77	0.83	9.2
0.2/0.9	1.86	0.86	10.0	1.85	0.84	11.7
0.3/0.5	2.06	0.87	6.7	1.77	0.84	7.8
0.3/0.7	2.12	0.87	8.1	1.84	0.84	9.5
0.3/0.9	2.09	0.87	**10.4**	1.67	0.82	**12.1**

Table 6: *Compression rates on the AVTSE task*

	K=128			K=64		
p_a/p_v	PESQ	STOI	r	PESQ	STOI	r
0.0/0.0	**2.60**	**0.92**	4.6	**2.39**	**0.91**	5.3
0.1/0.5	2.33	**0.92**	6.2	2.28	**0.91**	7.2
0.1/0.7	2.46	**0.92**	7.4	2.00	0.88	8.6
0.1/0.9	2.34	0.91	9.2	1.80	0.85	10.8
0.2/0.5	2.23	0.91	6.5	1.93	0.89	7.6
0.2/0.7	2.28	0.91	7.9	1.81	0.86	9.2
0.2/0.9	1.85	0.88	10.0	1.88	0.85	11.7
0.3/0.5	2.09	0.90	6.7	1.78	0.87	7.8
0.3/0.7	2.15	0.89	8.1	1.87	0.87	9.5
0.3/0.9	2.03	0.88	**10.4**	1.55	0.81	**12.1**

Following the approach in [18], we evaluate the inference latencies per frame on a system equipped with an NVIDIA RTX 2080 Ti GPU and an Intel(R) Xeon(R) Gold 5115 CPU, with results averaged over 100 utterances. The full-sized oTF-CrossNet and oTF-CrossNet-Mamba (B = 12) exhibit latencies of 5.84 ms and 4.43 ms, respectively. The causal visual encoder contributes approximately 0.3 ms, resulting in overall inference latencies of 6.14 ms for oAV-CrossNet and 4.73 ms for oAV-CrossNet-Mamba.

5. Conclusion

We have introduced online AV-CrossNet, a computationally efficient audiovisual system for causal and real-time AVSE and AVTSE. By integrating causal convolutional layers and leveraging model compression techniques, we substantially enhance the system's efficiency, achieving a minimal inference latency while maintaining competitive performance. Compared to the original AV-CrossNet, online AV-CrossNet operates frame-by-frame with a one-frame look-ahead and shrinks the model size up to 10 times. These findings underscore the potential of the proposed system for real-time AVSE applications in acoustically challenging environments.

6. Acknowledgements

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

7. References

[1] D. L. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, pp. 1702–1726, 2018.

[2] H. Wang and D. L. Wang, "Cross-domain diffusion based speech enhancement for very noisy speech," in *Proc. ICASSP*, 2023, pp. 1–5.

[3] Z. Zhao, D. Yang, R. Gu, H. Zhang, and Y. Zou, "Target confusion in end-to-end speaker extraction: Analysis and approaches," in *Proc. INTERSPEECH*, 2022, pp. 5333–5337.

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

[5] B. Rivet, W. Wang, S. M. Naqvi, and J. A. Chambers, "Audio-visual speech source separation: An overview of key methodologies," *IEEE Sig. Process. Mag.*, vol. 31, pp. 125–134, 2014.

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

[7] J. Wu, Y. Xu, S.-X. Zhang, L.-W. Chen, M. Yu, L. Xie, and D. Yu, "Time domain audio visual speech separation," in *Proc. ASRU*, 2019, pp. 667–673.

[8] R. Gao and K. Grauman, "VisualVoice: Audio-visual speech separation with cross-modal consistency," in *Proc. CVPR*, 2021, pp. 15 490–15 500.

[9] Z. Pan, R. Tao, C. Xu, and H. Li, "MuSE: Multi-modal target speaker extraction with visual cues," in *Proc. ICASSP*, 2021, pp. 6678–6682.

[10] J. Lin, X. Cai, H. Dinkel, J. Chen, Z. Yan, Y. Wang, J. Zhang, Z. Wu, Y. Wang, and H. Meng, "AV-SepFormer: Cross-attention sepformer for audio-visual target speaker extraction," in *Proc. ICASSP*, 2023, pp. 1–5.

[11] Z. Pan, G. Wichern, Y. Masuyama, F. G. Germain, S. Khurana, C. Hori, and J. Le Roux, "Scenario-aware audio-visual TF-GridNet for target speech extraction," in *Proc. ASRU*, 2023, pp. 1–8.

[12] A. L. A. Blanco, C. Valentini-Botinhao, O. Klejch, M. Gogate, K. Dashtipour, A. Hussain, and P. Bell, "AVSE Challenge: Audio-visual speech enhancement challenge," in *Proc. SLT Workshop*, 2023, pp. 465–471.

[13] M. Gogate, K. Dashtipour, A. Adeel, and A. Hussain, "Cochleanet: A robust language-independent audio-visual model for real-time speech enhancement," *Inf. Fusion*, vol. 63, pp. 273–285, 2020.

[14] M. Gogate, K. Dashtipour, and A. Hussain, "Towards real-time privacy-preserving audio-visual speech enhancement," in *Proc. SPSC*, 2022, pp. 7–10.

[15] Z. Zhu, H. Yang, M. Tang, Z. Yang, S. E. Eskimez, and H. Wang, "Real-time audio-visual end-to-end speech enhancement," in *Proc. ICASSP*, 2023, pp. 1–5.

[16] J. F. Montesinos, V. S. Kadandale, and G. Haro, "VoVit: Low latency graph-based audio-visual voice separation transformer," in *Proc. ECCV*, 2022, pp. 310–326.

[17] R. Gu, S.-X. Zhang, Y. Xu, L. Chen, Y. Zou, and D. Yu, "Multi-modal multi-channel target speech separation," *IEEE J. Sel. Top. Signal Process.*, vol. 14, pp. 530–541, 2020.

[18] H. Chen, R. Mira, S. Petridis, and M. Pantic, "RT-LA-VocE: Real-time low-SNR audio-visual speech enhancement," in *Proc. INTERSPEECH*, 2024, pp. 2215–2219.

[19] V. A. Kalkhorani, C. Yu, A. Kumar, K. Tan, B. Xu, and D. L. Wang, "AV-CrossNet: An audiovisual complex spectral mapping network for speech separation by leveraging narrow-and cross-band modeling," *arXiv:2406.11619*, 2024.

[20] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, pp. 8717–8727, 2018.

[21] V. A. Kalkhorani and D. Wang, "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.*, vol. 32, pp. 4999–5009, 2024.

[22] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, pp. 1256–1266, 2019.

[23] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," in *Proc. Speech Synth. Workshop*, 2016, p. 125.

[24] Y. Wu and K. He, "Group normalization," in *Proc. ECCV*, 2018, pp. 3–19.

[25] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," in *Proc. COLM*, 2024, pp. 1–32.

[26] C. Quan and X. Li, "Multichannel long-term streaming neural speech enhancement for static and moving speakers," *IEEE Signal Process. Lett.*, vol. 31, pp. 2295–2299, 2024.

[27] S. Han, H. Mao, and W. J. Dally, "Deep Compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," in *Proc. ICLR*, 2016, pp. 1–14.

[28] K. Tan and D. Wang, "Towards model compression for deep learning based speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1785–1794, 2021.

[29] Z. Pan, M. Ge, and H. Li, "A hybrid continuity loss to reduce over-suppression for time-domain target speaker extraction," in *Proc. INTERSPEECH*, 2022, pp. 1786–1790.

[30] T. Afouras, J. S. Chung, and A. Zisserman, "LRS3-TED: A large-scale dataset for visual speech recognition," *arXiv:1809.00496*, 2018.

[31] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *Proc. ICASSP*, 2017, pp. 776–780.

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

[33] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," in *Proc. Meet. Acoust.*, vol. 19, 2013, pp. 1–6.

[34] C. K. Reddy, H. Dubey, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "ICASSP 2021 deep noise suppression challenge," in *Proc. ICASSP*, 2021, pp. 6623–6627.

[35] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. INTERSPEECH*, 2018, pp. 1086–1090.

[36] N. Harte and E. Gillen, "TCD-TIMIT: An audio-visual corpus of continuous speech," *IEEE Trans. Multimed.*, vol. 17, pp. 603–615, 2015.

[37] V. A. Kalkhorani, A. Kumar, K. Tan, B. Xu, and D. Wang, "Time-domain transformer-based audiovisual speaker separation," in *Proc. INTERSPEECH*, 2023, pp. 3472–3476.

[38] D. Kim, S.-W. Chung, H. Han, Y. Ji, and H.-G. Kang, "HD-Demucs: General speech restoration with heterogeneous decoders," in *Proc. INTERSPEECH*, 2023, pp. 3829–3833.

[39] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. ICASSP*, 2020, pp. 46–50.