ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3315-6701-9/26/$31.00 ©2026 IEEE | DOI: 10.1109/ICASSP55912.2026.11460644

AUDIOVISUAL SPEECH ENHANCEMENT AND VOICE ACTIVITY DETECTION USING GENERATIVE AND SPEECH RECOGNITION FEATURES

Cheng Yu[1], *Vahid Ahmadi Kalkhorani*[1], *Anurag Kumar*[2], *Ke Tan*[2], *Buye Xu*[2], *DeLiang Wang*[3]

[1] Department of Computer Science and Engineering, The Ohio State University, USA
[2] Meta Reality Labs, USA
[3] School of Data Science, The Chinese University of Hong Kong, Shenzhen, China

ABSTRACT

In this study, we present an audiovisual speech enhancement (AVSE) system that extends AV-CrossNet by incorporating two stages of audiovisual fusion to address both speech enhancement (SE) and voice activity detection (VAD). The two-stage fusion consists of signal-level fusion, which uses generative lip-to-speech conversion from lip movements to provide noise-free time-frequency features, and embedding-level fusion, which integrates embeddings from an audio-visual speech recognition encoder into TF-CrossNet layers. Additionally, we introduce multi-task training to perform both SE and VAD with a VAD decoder. Specifically, our system is capable of jointly perform SE and VAD within a unified architecture. We evaluate the proposed system on multiple datasets, including COG-MHEAR challenge, LRS3-CHiME3, and LRS3-AudioSet. Experimental results show that our model achieves state-of-the-art SE performance across all datasets and superior VAD results compared to the audio-only baseline.

Index Terms— speech enhancement, audiovisual speech enhancement, voice activity detection, multi-modal

1. INTRODUCTION

Speech enhancement (SE) aims to enhance the quality and intelligibility of speech corrupted by noise, room reverberation, and other interferences [1]. Over the past few decades, SE algorithms have been dramatically advanced, especially since the introduction of deep learning techniques into this field [2]. Despite these achievements, SE algorithms still face limitations when applied to extremely noisy scenarios [3, 4]. Researchers have explored combining auditory and visual modalities to address these limitations [5]. These two modalities are complementary and compensate each other in human speech perception [6]. Specifically, the visual modality is unaffected by acoustic interference and can be directly utilized in voice activity detection (VAD) [7], speech recognition [8], and other applications [9]. Consequently, the visual modality can provide stable speech features in various acoustic scenarios, leading to the generalizability and robustness of AVSE [10, 11] in real-world environments where acoustic noise poses significant challenges.

There are various approaches to AVSE. Some algorithms utilize audio and video to produce Mel spectrograms [3], or speech codecs [12]. Others employ variational autoencoders [13, 14]. These methods have proven effective in generating quality speech signals under extremely noisy conditions, such as with multiple interferences and low signal-to-noise ratio (SNR) levels [3, 13]. However, the generated signals by these methods heavily depend on audio decoders and require a complex encoder-decoder training process. Furthermore, AVSE algorithms utilize visual features such as mouth images [10], face landmarks [15], or visual speech or speaker recognition embeddings [11, 16–18]. These approaches feature simpler regressive network architectures and more efficient training [19]. Among regressive AVSE algorithms, AV-CrossNet [17] achieves state-of-the-art performance in speaker separation (SS) and target speaker extraction (TSE). It uses pre-extracted audiovisual speech recognition (AVSR) features [20] through a linear fusion layer.

Recently, cross-modal generative approaches, specifically those converting video to audio, have garnered significant interest [21–23]. These approaches are fully insensitive to background interference and can be trained on large-scale datasets independent of downstream tasks. While they have a demonstrated capability in speech synthesis, the synthesized signals do not fully match natural speech at the signal level [23].

To leverage the advantages of the generative models, we propose a regressive model for AVSE that utilizes generative features. Specifically, our AVSE system incorporates both generative and AVSR features. We propose a novel two-stage fusion technique that incorporates two kinds of visual features. In signal-level fusion, we incorporate the Mel spectrograms from the diffusion-based generative LipVoicer model [23] and the complex spectrograms from input signals before feeding into a time-frequency (T-F) fusion encoder.

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).

Authorized licensed use limited to: The Ohio State University. Downloaded on June 08,2026 at 17:31:37 UTC from IEEE Xplore. Restrictions apply.

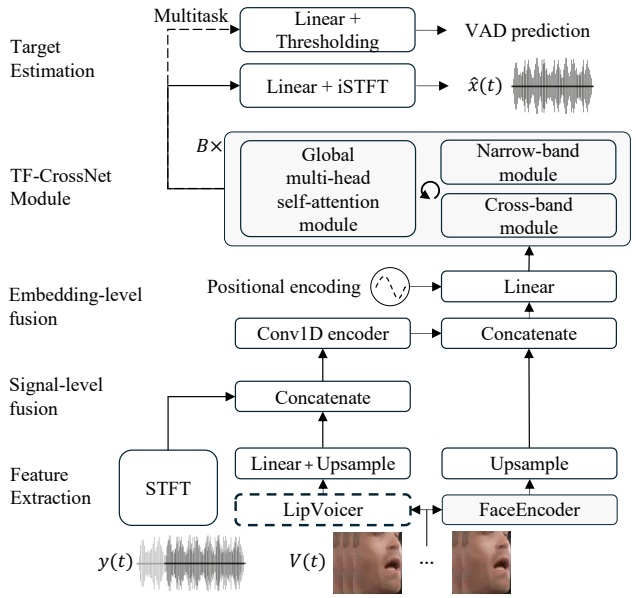

Fig. 1: Overall system architecture. Visual cues $\mathbf{V}$ are fed to LipVoicer and the DeepAVSR encoder to generate two levels of visual features for two-stage audiovisual fusion. The dash-boxed LipVoicer indicates its parameters are fixed during training and test.

In embedding-level fusion, we make use of a fine-tuned DeepAVSR [20] visual encoder. Our proposed model is designed to perform both SE and noise-robust VAD tasks. 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 the state-of-the-art performance on the second COG-MHEAR AVSE challenge, the LRS3-CHiME3, and the LRS3-AudioSet datasets, along with superior VAD results.

2. METHODOLOGY

Given a recorded visual stream $\mathbf{V}$ and a noisy speech signal of a single speaker $y(t)$ with voice activity of $d(t)$, we can formulate $y(t)$ in terms of the target speech signal $x(t)$ and additive noise signal $n(t)$ as

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

where t is the time index. Our proposed AVSE system takes $\mathbf{y}$ and $\mathbf{V}$ as inputs and estimates the target speech, $\tilde{\mathbf{x}}$, and voice activity, $\tilde{\mathbf{d}}$, as

$$\tilde{\mathbf{x}},\ \tilde{\mathbf{d}} = \text{AVSE}\ (\mathbf{y},\ \mathbf{V}). \tag{2}$$

Our proposed model is designed to strike a balance between generative and regressive-based visual encoders. To do so, we propose a two-stage audiovisual fusion that takes both generative and regressive features at two fusion stages.

2.1. Two-stage AV fusion

As illustrated in Fig. 1, the proposed two-stage audiovisual fusion module encodes the visual cues $\mathbf{V}$ with two encoders, then performs audiovisual fusion in two different levels. We employ LipVoicer [23] and FaceEncoder as visual encoders. LipVoicer [23] is a diffusion-based generative model that performs modality transformation. It transforms video into Mel spectrograms and is capable of reconstructing speech from silent lip movements. This encoder is fixed during training and test. FaceEncoder is a visual encoder employed for AVSR task in DeepAVSR [20]. This encoder is fine-tuned during training.

The two-stage audiovisual fusion can be formulated as follows. First, the signal-level fusion can be formulated as

$$\mathbf{Y_r}, \mathbf{Y_i} = \text{STFT}(\mathbf{y}), \quad \mathbf{Y_r}, \mathbf{Y_i} \in \mathbb{R}^{T\times 1\times F}, \tag{3}$$

$$\mathbf{M} = \text{LipVoicer}(\mathbf{V}), \quad \mathbf{M} \in \mathbb{R}^{T_L\times 1\times F_L}, \tag{4}$$

$$\mathbf{M'} = \text{Linear}(\text{Upsample}(\mathbf{M})), \quad \mathbf{M'} \in \mathbb{R}^{T\times 1\times F}, \tag{5}$$

$$\mathbf{Y'} = \text{Concatenate}(\mathbf{Y_r}, \mathbf{Y_i}, \mathbf{M'}), \quad \mathbf{Y'} \in \mathbb{R}^{T\times 3\times F}, \tag{6}$$

$$\mathbf{E1} = \text{T-F Encoder}(\mathbf{Y'}), \quad \mathbf{E1} \in \mathbb{R}^{T\times C\times F}. \tag{7}$$

Through fusing the noise-free T-F feature $\mathbf{M'}$ with the complex spectrogram $[\mathbf{Y_r}, \mathbf{Y_i}]$, we aim to achieve a more robust complex spectral mapping. Second, the embedding-level fusion can be formulated as

$$\mathbf{V'} = \text{FaceEncoder}(\mathbf{V}), \quad \mathbf{V} \in \mathbb{R}^{T_v\times C\times F}, \tag{8}$$

$$\mathbf{E2} = \text{Upsample}(\mathbf{V'}), \quad \mathbf{E2} \in \mathbb{R}^{T\times C\times F}, \tag{9}$$

$$\mathbf{E3} = \text{Concatenate}(\mathbf{E1}, \mathbf{E2}), \quad \mathbf{E3} \in \mathbb{R}^{T\times 2C\times F}, \tag{10}$$

$$\mathbf{E} = \text{Linear}(\mathbf{E3}), \quad \mathbf{E} \in \mathbb{R}^{T\times C\times F}. \tag{11}$$

The high-dimensional feature $\mathbf{E}$ is then fed to the TF-CrossNet blocks after adding the random chunk positional encoding [24].

2.2. Multi-task (SE and VAD)

As illustrated in Fig. 1, the proposed AVSE system can be trained either for speech enhancement alone or as a multi-task model that handles both SE and VAD. The multi-task variant only requires an additional simple linear decoder. In this model, the features from the last TF-CrossNet layer are fed to the SE and VAD decoders. The VAD decoder outputs $\tilde{\mathbf{d}}$, which contains one-hot labels for each time frame. We apply the signal-processing-based rVAD [25] algorithm to clean utterances to generate the target VAD labels.

3. EXPERIMENTAL SETUP

3.1. Datasets

In this paper, we use the second COG-MHEAR challenge dataset [26] and LRS3-CHiME3 dataset [27] to evaluate our

Authorized licensed use limited to: The Ohio State University. Downloaded on June 08,2026 at 17:31:37 UTC from IEEE Xplore. Restrictions apply.

model on the AVSE task. Furthermore, we use the LRS3-AudioSet dataset to assess the performance of the proposed method on joint AVSE and audiovisualVAD tasks. From the COG-MHEAR challenge training set, we prepare the validation dataset using 24 speakers with the largest number of utterances, resulting in 1339 utterances. The rest of the utterances in the training set serve as our training dataset. We treat the SE utterances in the development set as the test dataset. The training, validation, and test datasets contain 30,297, 1,339, and 1,661 utterances. The utterance length for both training and validation datasets is 3.0 seconds. The test dataset contains utterances of around 2 seconds to 30 seconds. All utterances in these datasets utilize speech-weighted SNRs [26], with SNR values randomly drawn from [-10, 10] dB range. For the LRS3-CHiME3 dataset, we follow the instructions in [28] to prepare the test dataset. We select 244 utterances with durations no shorter than 3.2 seconds and randomly mix these utterances with noise signals from the CHiME3 dataset, using the standard SNR drawn from [-6, 12] dB range. For the LRS3-AudioSet dataset, we prepare the training, validation, and test dataset using the train-val set from the original LRS3 dataset. We create the training, validation, and test datasets consisting of around 100k, 10k, and 2.1k utterances. Moreover, we select non-speech noise signals from the AudioSet, resulting in 2.7k, 305, and 86 signals for training, validation, and test datasets. We randomly mix these utterances and non-speech signals using the standard SNR selected from [-15, 15] dB range. For VAD experiments, we create more challenging mixtures with the SNR range of [-35, 15] dB to present the effectiveness of our proposed model.

3.2. Baselines

We prepared several audio-only (AO) and audiovisual models to serve as baselines. For AO models, we trained the TF-CrossNet [24] and the HD-DEMUCS [29] to validate the effectiveness of the visual modality. Note that the HD-DEMUCS is the current best variant of the DEMUCS [30] for SE. Furthermore, to verify the effectiveness of the proposed signal-level fusion, we trained the AV-HD-DEMUCS using the signal-level fusion. Also, we prepared AVDPRNN [16] and AVGridNet [16] following the instructions from the original paper and the publicly accessible codes[1]. These two models utilize the pre-extracted DeepAVSR [20] features through an early-fusion mechanism. Last but not least, we prepared a multi-task TF-CrossNet to serve as an AO multi-task baseline.

3.3. Loss function

In this paper, we train all SE models with hybrid loss [31] for fair comparison. The hybrid loss consists of scale-invariant signal-to-distortion ratio (SI-SDR) loss and a multi-resolution short-time Fourier transform (STFT) loss. We adopt a standard SI-SDR loss that scales the target signal to match the estimated signal. For STFT loss, we adopt STFT windows with lengths of 256, 512, and 1024, and shifts of 25, 60, and 120. Furthermore, we add binary cross entropy (BCE) loss to the hybrid loss to train the multi-task model.

3.4. Evaluation Metrics

To evaluate the quality and intelligibility of the estimated speech, we adopt perceptual evaluation of speech quality (PESQ), short-time objective intelligibility (STOI), and SI-SDR metrics from the Torch-Metrics [32] package. Moreover, to evaluate the effectiveness of VAD, we adopt the Hit-FA [33] rate. The Hit-FA is composed of the Hit rate (True positive rate) and the FA rate (False positive rate). According to [33], the Hit-FA is a more effective way to reveal the performance of a VAD algorithm.

3.5. Hyperparameters

All utterances are resampled to 16 kHz. All videos are sampled with 25 frames per second (FPS). We adopt STFT with window/shift sizes of 256/128 for audio signals, resulting in 129 frequency bins in each feature frame. The LipVoicer produces a Mel spectrogram of 80 frequency bins with a 160 shift size. The DeepAVSR encoder produces a 512-dimensional feature with a 640 shift size. We apply temporal linear interpolation to upsample these features, aligning them with the audio feature frame rate. Additionally, we apply linear layers to the video features to align their feature dimensions with those of the audio features' time-frequency resolution. We use 12 TF-CrossNet modules ($B = 12$) in the model architecture. For the training of all models, we adopt the ReduceLROnPlateau learning rate scheduler. We set the decay factor at 0.85 and the patience of 3 epochs. For the proposed system, the FaceEncoder is initialized with pre-trained DeepAVSR weights and fine-tuned during training, whereas the LipVoicer encoder is kept fixed.

4. EXPERIMENTAL RESULTS

4.1. COG-MHEAR AVSE challenge

Table 1 presents the results on the second COG-MHEAR challenge dataset. The proposed model outperforms all other approaches across every metric, achieving a 0.11 PESQ gain over the audio-only TF-CrossNet. Compared to AV-CrossNet, it further improves by 0.05 and 0.2 in PESQ and SI-SDR, respectively, through its advanced two-stage audio-visual (AV) fusion. While AV-CrossNet improved over TF-CrossNet by 0.06 in PESQ, our model delivers an additional 0.05 gain, underscoring the significance of the improvement over AV-CrossNet.

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

Authorized licensed use limited to: The Ohio State University. Downloaded on June 08,2026 at 17:31:37 UTC from IEEE Xplore. Restrictions apply.

Table 1: Results on the COG-MHEAR AVSE Challenge dataset.

Method	AV	PESQ	STOI	SI-SDR
Unprocessed	-	1.15	0.68	-4.4
AV-DPRNN [16]	✓	2.02	0.86	11.4
AV-GridNet [16]	✓	2.68	0.91	14.2
TF-CrossNet [24]	✗	2.69	0.91	14.3
AV-CrossNet [17]	✓	2.75	**0.92**	14.3
Proposed	✓	**2.80**	**0.92**	**14.5**

4.2. LRS3-CHiME3

In Table 2, the models in the lower section are trained on the LRS3-AudioSet dataset and evaluated on the mismatched LRS3-CHiME3 dataset. The proposed model outperforms all competing approaches across every evaluation metric, with particularly notable gains in PESQ (at least 0.05 over AV-GridNet and 0.5 over the remaining models). While AV-GridNet delivers competitive results, it is considerably less efficient [34]. We also highlight that AV-HD-DEMUCS surpasses its AO counterpart by 0.16 in PESQ. Since AV-HD-DEMUCS incorporates only signal-level fusion, this result provides clear evidence of the effectiveness of such fusion.

Table 2: Results on the LRS3-CHiME3 dataset.

Method	AV	PESQ	ESTOI	SI-SDR
Mixture	-	1.20	0.77	2.6
SGMSE+ [28] [27]	✗	2.08	0.89	10.6
VisualVoice [27] [35]	✓	2.11	0.89	11.0
AV-Gen [27]	✓	2.23	0.90	11.4
HD-DEMUCS [29]	✗	2.11	0.89	11.9
AV-HD-DEMUCS [29]	✓	2.27	0.90	12.5
AV-DPRNN [16]	✓	2.34	0.92	13.6
AV-GridNet [16]	✓	2.79	**0.94**	14.0
Proposed	✓	**2.84**	**0.94**	**14.1**

4.3. LRS3-AudioSet

Table 3 shows the performances on LRS3-AudioSet. The proposed model outperforms other baselines by a significant margin in PESQ (at least 0.07) and shows superior performances in STOI and SI-SDR. Furthermore, the AV-HD-DEMUCS and the proposed model outperform their AO counterparts in all metrics, indicating the effectiveness of the signal-level fusion and the two-stage fusion, respectively.

In Fig. 2, we compare our proposed model and TF-CrossNet with their VAD results. Our proposed AVSE model slightly outperforms the TF-CrossNet* in positive SNR ranges, where the target speech dominates the mixture. However, the performance gap between the proposed model and TF-CrossNet* gradually grows bigger when tested in lower SNR ranges. This indicates that our proposed model can provide a noise-robust VAD even in extremely challenging conditions. We also observe relatively high Hit-FA rates for both methods in the higher SNR ranges, which we attribute to the high speech proportion (approximately 85%) in the LRS3-AudioSet dataset.

Table 3: Results on the LRS3-AudioSet dataset.

Method	AV	PESQ	STOI	SI-SDR
Mixture	-	1.32	0.72	0.08
HD-DEMUCS [29]	✗	2.25	0.83	10.7
AV-HD-DEMUCS [29]	✓	2.35	0.87	11.8
AV-DPRNN [16]	✓	2.33	0.88	12.4
TF-CrossNet* [24]	✗	2.79	0.90	12.6
AV-GridNet [16]	✓	2.81	**0.92**	13.2
Proposed*	✓	**2.88**	**0.92**	**13.6**

* Indicates the multi-task configuration.

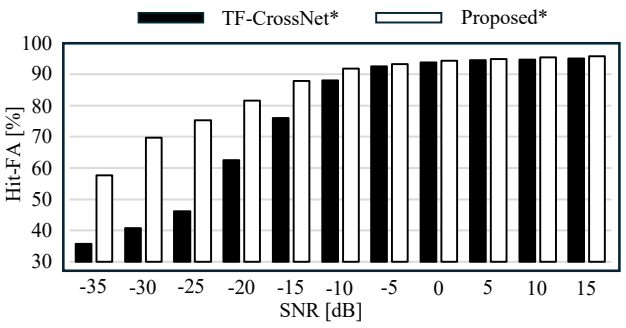

Fig. 2: Hit-FA results of the proposed model across different SNRs for unprocessed mixtures on the test dataset of LRS3-AudioSet.

5. CONCLUSION

In this work, we present an AVSE system with a novel two-stage audiovisual fusion framework. The fusion combines generative features from LipVoicer with visual speech recognition embeddings from DeepAVSR. While LipVoicer provides noise-free time–frequency features to enhance complex spectral mapping, the DeepAVSR encoder is fine-tuned to jointly support SE and VAD tasks. Experimental results demonstrate that the proposed system achieves state-of-the-art performance across multiple datasets and delivers superior VAD results compared to its AO counterpart.

Authorized licensed use limited to: The Ohio State University. Downloaded on June 08,2026 at 17:31:37 UTC from IEEE Xplore. Restrictions apply.

6. REFERENCES

[1] P. C. Loizou, "Speech enhancement: Theory and practice," *CRC press*, 2013.

[2] De Liang Wang and Jitong Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, pp. 1702–1726, 2018.

[3] Rodrigo Mira, Buye Xu, Jacob Donley, Anurag Kumar, Stavros Petridis, Vamsi Krishna Ithapu, and Maja Pantic, "LA-VocE: Low-SNR audio-visual speech enhancement using neural vocoders," in *Proc. ICASSP*. IEEE, 2023, pp. 1–5.

[4] Heming Wang and De Liang Wang, "Cross-domain diffusion based speech enhancement for very noisy speech," in *Proc. ICASSP*. IEEE, 2023, pp. 1–5.

[5] Daniel Michelsanti, Zheng-Hua Tan, Shi-Xiong Zhang, Yong Xu, Meng Yu, Dong Yu, and Jesper Jensen, "An overview of deep-learning-based audio-visual speech enhancement and separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1368–1396, 2021.

[6] Jean-Luc Schwartz, Frédéric Berthommier, and Christophe Savariaux, "Seeing to hear better: evidence for early audio-visual interactions in speech identification," *Cognition*, vol. 93, pp. B69–B78, 2004.

[7] Muhammad Shahid, Cigdem Beyan, and Vittorio Murino, "S-VVAD: Visual voice activity detection by motion segmentation," in *Proc. Winter Conference on Applications of Computer Vision*, 2021, pp. 2332–2341.

[8] Pingchuan Ma, Stavros Petridis, and Maja Pantic, "Visual speech recognition for multiple languages in the wild," *Nature Machine Intelligence*, vol. 4, pp. 930–939, 2022.

[9] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," in *Proc. Multimedia*, 2020, pp. 484–492.

[10] Jen-Cheng Hou, Syu-Siang Wang, Ying-Hui Lai, Yu Tsao, Hsiu-Wen Chang, and Hsin-Min Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, pp. 117–128, 2018.

[11] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, "The Conversation: deep audio-visual speech enhancement," *Proc. INTERSPEECH*, pp. 3244–3248, 2018.

[12] Karren Yang, Dejan Marković, Steven Krenn, Vasu Agrawal, and Alexander Richard, "Audio-visual speech codecs: Rethinking audio-visual speech enhancement by re-synthesis," in *Proc. CVPR*, 2022, pp. 8227–8237.

[13] Mostafa Sadeghi, Simon Leglaive, Xavier Alameda-Pineda, Laurent Girin, and Radu Horaud, "Audio-visual speech enhancement using conditional variational auto-encoders," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1788–1800, 2020.

[14] Mostafa Sadeghi and Xavier Alameda-Pineda, "Mixture of inference networks for VAE-based audio-visual speech enhancement," *IEEE Transactions on Signal Processing*, vol. 69, pp. 1899–1909, 2021.

[15] Giovanni Morrone, Sonia Bergamaschi, Luca Pasa, Luciano Fadiga, Vadim Tikhanoff, and Leonardo Badino, "Face landmark-based speaker-independent audio-visual speech enhancement in multi-talker environments," in *Proc. ICASSP*. IEEE, 2019, pp. 6900–6904.

[16] Zexu Pan, Gordon Wichern, Yoshiki Masuyama, François G Germain, Sameer Khurana, Chiori Hori, and Jonathan Le Roux, "Scenario-aware audio-visual TF-GridNet for target speech extraction," in *Proc. ASRU*. IEEE, 2023, pp. 1–8.

[17] Vahid Ahmadi Kalkhorani, Cheng Yu, Anurag Kumar, Ke Tan, Buye Xu, and De Liang Wang, "AV-CrossNet: an audiovisual complex spectral mapping network for speech separation by leveraging narrow-and cross-band modeling," *arXiv:2406.11619*, 2024.

[18] Zhongbo Sun, Yannan Wang, and Li Cao, "An attention based speaker-independent audio-visual deep learning model for speech enhancement," in *Proc. MMM*. Springer, 2020, pp. 722–728.

[19] Zirun Zhu, Hemin Yang, Min Tang, Ziyi Yang, Sefik Emre Eskimez, and Huaming Wang, "Real-time audio-visual end-to-end speech enhancement," in *Proc. ICASSP*. IEEE, 2023, pp. 1–5.

[20] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman, "Deep audio-visual speech recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, pp. 8717–8727, 2018.

[21] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar, "Learning individual speaking styles for accurate lip to speech synthesis," in *Proc. CVPR*. 2020, IEEE.

[22] Wei-Ning Hsu, Tal Remez, Bowen Shi, Jacob Donley, and Yossi Adi, "ReVISE: Self-Supervised Speech Resynthesis with Visual Input for Universal and Generalized Speech Regeneration," in *Proc. CVPR*. 2023, pp. 18795–18805, IEEE.

[23] Yochai Yemini, Aviv Shamsian, Lior Bracha, Sharon Gannot, and Ethan Fetaya, "LipVoicer: generating speech from silent videos guided by lip reading," in *Proc. ICLR*, 2024, pp. 1–20.

[24] Vahid Ahmadi Kalkhorani and DeLiang Wang, "TF-CrossNet: Leveraging Global, Cross-Band, Narrow-Band, and Positional Encoding for Single-and Multi-Channel Speaker Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[25] Zheng-Hua Tan, Achintya kr. Sarkar, and Najim Dehak, "rVAD: an unsupervised segment-based robust voice activity detection method," *Computer Speech & Language*, vol. 59, pp. 1–21, 2020.

[26] Andrea Lorena Aldana Blanco, Cassia Valentini-Botinhao, Ondrej Klejch, Mandar Gogate, Kia Dashtipour, Amir Hussain, and Peter Bell, "AVSE Challenge: audio-visual speech enhancement challenge," in *Proc. SLT Workshop*. 2023, pp. 465–471, IEEE.

[27] Julius Richter, Simone Frintrop, and Timo Gerkmann, "Audio-Visual Speech Enhancement with Score-Based Generative Models," in *Proc. ITG Speech Communication*. VDE, 2023, pp. 275–279.

[28] Julius Richter, Simon Welker, Jean-Marie Lemercier, Bunlong Lay, and Timo Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2351–2364, 2023.

[29] Doyeon Kim, Soo-Whan Chung, Hyewon Han, Youna Ji, and Hong-Goo Kang, "HD-Demucs: General speech restoration with heterogeneous decoders," in *Proc. INTERSPEECH*, 2023, pp. 3829–3833.

[30] Alexandre Defossez, Gabriel Synnaeve, and Yossi Adi, "Real Time Speech Enhancement in the Waveform Domain," in *Proc. INTERSPEECH*, 2020, pp. 3291–3295.

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

[32] Nicki Skafte Detlefsen, Jiri Borovec, Justus Schock, Ananya Harsh Jha, Teddy Koker, Luca Di Liello, Daniel Stancl, Changsheng Quan, Maxim Grechkin, and William Falcon, "TorchMetrics-measuring reproducibility in pytorch," *Journal of Open Source Software*, vol. 7, pp. 4101, 2022.

[33] Xiao-Lei Zhang and De Liang Wang, "Boosting contextual information for deep neural network based voice activity detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, pp. 252–264, 2015.

[34] Vahid Ahmadi Kalkhorani and De Liang Wang, "CrossNet: Leveraging global, cross-band, narrow-band, and positional encoding for single- and multi-channel speaker separation," *arXiv: 2403.03411*, 2024.

[35] Ruohan Gao and Kristen Grauman, "VisualVoice: audio-visual speech separation with cross-modal consistency," in *Proc. CVPR*. IEEE, 2021, pp. 15490–15500.

Authorized licensed use limited to: The Ohio State University. Downloaded on June 08,2026 at 17:31:37 UTC from IEEE Xplore. Restrictions apply.