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

ClearerVoice-Studio: Bridging Advanced Speech Processing Research and Practical Deployment

Shengkui Zhao, Zexu Pan, Bin Ma

[1]Tongyi Lab, Alibaba Group, Singapore

{shengkui.zhao, zexu.pan, b.ma}@alibaba-inc.com

Abstract

This paper introduces ClearerVoice-Studio, an open-source, AI-powered speech processing toolkit designed to bridge cutting-edge research and practical application. Unlike broad platforms like SpeechBrain and ESPnet, ClearerVoice-Studio focuses on interconnected speech tasks of speech enhancement, separation, super-resolution, and multimodal target speaker extraction. A key advantage is its state-of-the-art pretrained models, including FRCRN (3M+ uses) and MossFormer (2.5M+ uses), optimized for real-world scenarios. It also offers model optimization tools, multi-format audio support, the SpeechScore evaluation toolkit, and user-friendly interfaces, catering to researchers, developers, and end-users. Its rapid adoption (2.8K GitHub stars, 200+ forks) highlights its academic and industrial impact. This paper details ClearerVoice-Studio's [1] capabilities, architectures, training strategies, benchmarks, community impact, and future plan.

Index Terms: speech enhancement, speech separation, speech super-resolution, multi-modal target-speaker extraction, advanced speech processing

1. Introduction

Speech processing is now fundamental to modern technology, powering advancements in communication, voice interfaces, and multimedia [1]. While crucial for these applications, accurately processing speech is challenged by the often degraded quality of real-world audio. Noise, interfering speech, reverberation, and low resolution commonly corrupt recordings, severely impacting the performance of speech-based systems and human perception. Robust speech processing techniques are therefore essential for reliable operation in diverse and challenging environments [2, 3].

In recent years, robust speech processing has witnessed significant advancements, driven by cutting-edge research and innovative techniques. To democratize access to these advancements and foster collaborative innovation, several general-purpose speech open-source toolkits, such as SpeechBrain [4] and ESPnet [5], have emerged. These toolkits have accelerated research progress and provided essential resources for a wide range of speech-related tasks. However, they primarily focus on replicating results from research papers using public datasets, which often do not translate seamlessly to real-world deployment. In addition, there is a growing trend in multimodal speech processing, where modalities such as visual, text, or brain signals provide valuable complementary information [6]. However, general toolkits like SpeechBrain and ESPnet do not yet support these multimodal approaches. Bridging these gaps requires addressing several challenges, including a unified platform design for multimodal support, the need for large-scale real-world datasets, efficient model optimization, comprehensive performance evaluations, and user-friendly interfaces, all of which are essential for enabling practical deployment and broader adoption among end users.

In addition to general-purpose toolkits, numerous open-source repositories have focused on specific speech processing tasks. For example, Asteroid [7] specializes in audio source separation but falls short in model performance. DeepFilterNet [8] offers a lightweight speech enhancement solution but lacks broader functionality. Resemble Enhance[2] targets denoising for speech quality improvement using a dual-module approach. AudioSR [9] excels in audio super-resolution but struggles in noisy environments. FlowAVSE [10] specializes in audio-visual speech enhancement, particularly for real-time applications. While these tools offer valuable contributions for research or single-task solutions, they often lack the comprehensive scope and integration needed to address the complex, interconnected challenges present in real-world scenarios. For example, a noisy recording may also contain overlapping speakers, or a low-resolution recording might be further compromised by noise. As a result, there is a strong demand for a unified toolkit capable of tackling multiple challenges within a cohesive framework. Moreover, there is a need for a unified platform that enables benchmarking of multimodal speech processing methods within the research community, facilitating the development of better models.

To bridge these gaps, we introduce ClearerVoice-Studio, a versatile and user-friendly speech processing toolkit that leverages state-of-the-art techniques to address a wide range of real-world challenges. It seamlessly integrates solutions for speech enhancement (SE), speech separation (SS), super-resolution (SR), and multimodal target speaker extraction (AVSE) into a unified platform. By combining cutting-edge models, practical optimization and fine-tuning tools, comprehensive speech evaluation metrics, and an intuitive user-friendly interface, ClearerVoice-Studio empowers researchers, developers, and end-users to efficiently tackle complex speech processing tasks with ease and efficiency.

2. Functionalities and architectures

This section details ClearerVoice-Studio's functionalities across its target tasks and provides an in-depth look at the core model architectures that power these functionalities.

[1]https://github.com/modelscope/ClearerVoice-Studio

[2]https://github.com/resemble-ai/resemble-enhance

10.21437/Interspeech.2025-680

2.1. ClearerVoice-Studio functionalities

2.1.1. Easy access to pre-trained models

ClearerVoice-Studio provides a suite of advanced, pre-trained models designed for a variety of SE, SS, SR, and AVSE tasks.

SE task: ClearerVoice-Studio utilizes state-of-the-art architectures like FRCRN [11] and MossFormer2 [12] to deliver advanced noise reduction for speech signals. It includes three models: FRCRN_SE_16K and MossFormerGAN_SE_16K for 16 kHz audio, and MossFormer2_SE_48K for 48 kHz audio. The models process noisy audio inputs, whether in raw wave format or in common formats like MP3, OGG, and AAC, producing clean, denoised outputs while supporting multiple input sampling rates. By effectively mitigating ambient noise, interference, and reverberation, these models enhance speech clarity while preserving the natural characteristics of the original audio. ClearerVoice-Studio's speech enhancement capabilities are applicable across various scenarios, including improving the quality of meeting recordings, enhancing podcasts and audiobooks, optimizing online communication, pre-processing audio for voice cloning, conversion, or synthesis, and restoring historical recordings.

SS task: ClearerVoice-Studio offers MossFormer2_SS_16K for 16 kHz robust speech separation capabilities for scenarios involving overlapping speakers. Leveraging advanced architectures like MossFormer [13] and MossFormer2 [12], the toolkit effectively decomposes mixed audio signals into distinct, speaker-specific streams. Specifically, given a single-channel mixture containing multiple speakers, the output comprises individually separated signals for each. This functionality is crucial for applications like automatic transcription, speaker diarization, and multi-party communication systems, where accurately distinguishing and isolating individual speakers significantly improves downstream processing and overall speech intelligibility. By addressing the complexities of real-world audio environments, ClearerVoice-Studio ensures robust performance even in scenarios with high levels of overlapping speech and background noise.

SR task: ClearerVoice-Studio also provides MossFormer2_SR_48K, which upscales speech audio above 16 kHz to 48 kHz for enhancing low-resolution speech by restoring lost frequency components [14]. Trained on paired low- and high-resolution speech datasets, the speech super-resolution models leverage a combination architecture of MossFormer2 and HiFiGAN [15] to accurately reconstruct missing spectral details. Input as a low-resolution waveform yields an enhanced, high-resolution output with improved intelligibility and naturalness. This capability is particularly valuable for restoring degraded or compressed recordings, enhancing telecommunication audio, and improving speech quality in bandwidth-limited environments.

AVSE task: ClearerVoice-Studio further advances multimodal target speaker extraction by supporting a variety of conditioning modalities, including face-conditioned (AV-TFGridNet [16]), gesture-conditioned (SEG [17]), speech-conditioned (SpEx+ [18]), and brain-EEG-conditioned (NeuroHeed [19]) speaker extraction. The platform provides replicated pre-trained models from respective state-of-the-art research papers, alongside a newly proposed face-conditioned speaker extraction model named AV_MossFormer2_TSE_16K. This new model, trained on extensive datasets with advanced architectures, supports 16 kHz speaker isolation using synchronized face recordings. It can accurately extract a target speaker's voice, even in challenging conditions with overlapping speech or degraded audio quality.

These four functionalities, while available as individual modules, can be combined into composite processing pipelines. For example, speech enhancement can precede super-resolution to address background noise before spectral detail restoration. Similarly, super-resolution can be applied to the output of speech separation or multimodal target speaker extraction for further quality enhancement.

2.1.2. Comprehensive training scripts

ClearerVoice-Studio offers comprehensive training and fine-tuning scripts for its advanced speech processing suite, encompassing speech enhancement, separation, super-resolution, and multimodal target speaker extraction. Users can easily configure hyperparameters, loss functions, and evaluation metrics for task-specific model adaptation, significantly reducing development time and resource requirements.

2.1.3. Speech evaluation resource: SpeechScore

ClearerVoice-Studio includes SpeechScore, a comprehensive resource for evaluating the quality and intelligibility of processed speech. SpeechScore employs both subjective and objective evaluation methods, incorporating a range of common metrics (e.g., DNSMOS, BSSEval, PESQ, STOI, SRMR, MCD) to provide detailed insights into voice processing algorithm performance, ensuring high standards for clarity, naturalness, and overall quality.

2.1.4. User-friendly interfaces

ClearerVoice-Studio offers user-friendly interfaces (CLI and GUIs) for diverse users. The CLI caters to researchers and developers needing script-based interactions, providing flexible training and inference scripts. Web-based GUIs (HuggingFace[3] and ModelScope[4] Spaces) allow non-technical users to easily upload audio and apply models with real-time feedback. This dual approach promotes broader adoption in academic and industrial settings.

2.2. Core model architectures

ClearerVoice-Studio employs two core architectures: FRCRN and MossFormer2. FRCRN_SE_16K uses the FRCRN architecture, while all other models are based on MossFormer2.

2.2.1. FRCRN architecture

FRCRN [11] employs a convolutional-recurrent network (CRN) design, integrating a convolutional encoder-decoder (CED) with a recurrent structure. It uses complex-valued CNNs to capture local spectral patterns in spectrograms and complex-valued RNNs (FSMNs) for temporal dependency modeling. A key feature is its frequency recurrence mechanism, which processes frequency bands sequentially for fine-grained spectral detail capture. Operating on time-frequency spectrograms, FRCRN predicts the complex Ideal Ratio Mask (cIRM) and has achieved strong speech enhancement results, including notable performance in the ICASSP 2022 DNS Challenge. FRCRN_SE_16K adopts the FRCRN architecture with parameters matching the original paper [11], but concatenates two FRCRN networks for enhanced mask prediction.

[3] https://huggingface.co/spaces/alibabasglab/ClearVoice
[4] https://modelscope.cn/studios/iic/ClearerVoice-Studio

2.2.2. MossFormer2 architecture

MossFormer2 [12] employs a hybrid design that extends the Transformer mechanism proposed in MossFormer [13] by integrating a recurrent module. The Transformer mechanism strengths the modeling capabilities through a novel gated single-head Transformer with convolution-augmented joint self-attentions. Compared to traditional multi-head attentions, the gated single-head Transformer is able to offer more efficient handling of both long-range interactions and local features. In addition, the joint local and global self-attention addresses long-sequence modeling by performing full self-attention on local blocks and linearized self-attention across the entire sequence. Besides the capabilities provided by the Transformer mechanism, the recurrent module is stacked to the attention block to model intricate temporal dependencies within speech signals.

2.2.3. Adopting MossFormer2 across tasks

MossFormer2 serves as the primary feature mapping network architecture in ClearerVoice-Studio. MossFormer2_SE_48K uses MossFormer2 to predict the phase-sensitive mask (PSM) from mel-spectrograms derived from input waveforms, applying the mask to input complex spectrograms for output. MossFormerGAN_SE_16K applies MossFormer2 to 3D feature maps extracted by a time-frequency (T-F) encoder, performing simultaneous temporal and frequency attention. A dual-decoder structure (T-F masking and spectral decoders) predicts the complex mask and spectrum, respectively, combining them via weighted sum for the final output. MossFormer2_SR_48K employs a transformer-convolutional generator, combining MossFormer2 for enriched mel-spectrogram representation and HiFiGAN for upsampling and waveform reconstruction [14]. MossFormer2_SS_16K uses a time-domain encoder-separator-decoder framework, with MossFormer2 as the masking network predicting time-domain masks and 1D convolutional encoder and decoder handling feature extraction and waveform reconstruction, respectively. AV_MossFormer2_TSE_16K is an audio-visual input extension of MossFormer2_SS_16K. It repeatedly fuses visual features from an additional visual encoder with speech features at the beginning of the MossFormer-Recurrent module through frame-level concatenation and projection.

3. Training strategies and performance

3.1. Training strategies

3.1.1. Dataset preparation

ClearerVoice-Studio's pre-trained models are trained on a large and diverse dataset combining public and internal resources. For speech enhancement, the fullband clean speech sources include the 4th DNS-Challenge speech dataset [20] and the internal TTS dataset. The 4th DNS-Challenge provides clean speech of multiple languages mainly derived from Librivox [5] and VCTK corpus [21]. The audio clips were filtered based on DNS-MOS scores [20] provided by the DNS challenge organizer to control the speech quality. The VCTK corpus consists of 110 English speakers of 400 utterances each. The internal TTS dataset include 200 Mandarin speakers of 500 utterances each and 2 English speakers of 50000 utterances each. The total length of the fullband clean speech is 850 hours.

Full-band noise sources include the 4th DNS Challenge noise dataset, the WHAM!48kHz noise dataset[6], and internal recordings. The 4th DNS Challenge noise dataset is derived primarily from AudioSet [22], Freesound[7], and the DEMAND dataset [23]. AudioSet noise clips were filtered using a voice activity detector to remove speech presence. The WHAM!48kHz noise dataset consists mainly of recordings from restaurants, cafes, bars, and parks, while our internal recordings cover office and meeting room environments. The combined noise data totals 300 hours.

Our reverberant speech was generated using 100,000 synthetic room impulse responses based on RIR_generator[8] covering various room sizes. Noisy-clean data was created by mixing noise with clean or reverberant speech (with 7:3 ratio) at random SNRs between 0 and 15 dB. The resulting 8000-hour training and 50-hour development sets were augmented by downsampling 10% to 16 kHz and 5% to 8 kHz and then upsampling to 48 kHz. For 16 kHz wideband enhancement, the full-band data was resampled, supplemented by 262 hours of LibriTTS [27]. 48 kHz super-resolution used the full-band clean speech, with training sets created by low-pass filtering to 16-32 kHz. Speech separation used 107 hours of data from 110 VTCK, 202 internal TTS, and 1300 LibriTTS speakers, mixing two speakers' utterances at 0-5 dB SNR, and noise at -5 to 15 dB SNR relative to the loudest speaker. The full-band noise and impulse responses were downsampled and reused. For AV_MossFormer2_TSE_16K model, we adopt the Grid, TCD-Timit, LRS2, LRS3, VoxCeleb2, MISP as speech source dataset to simulate the speech mixtures, and uses FSD50K, WHAM noise, MUSAN noise for noise sources.

3.1.2. Model optimization

We employed several key training strategies [28] to ensure robust and efficient model development in ClearerVoice-Studio. Scalable distributed training with NCCL is used to enable multi-GPU support. Reproducibility is maintained through consistent random seeds and cuDNN settings. To facilitate multi-model

Table 1: *Speech enhancement DNS-2020 benchmark (16 kHz). The test set contains 300 synthetic clips without reverb. SNR levels are uniformly distributed between 0 dB and 25 dB.*

Model	PESQ	NB_PESQ	STOI	P808 MOS
Noisy	1.58	2.45	91.52	3.15
DCCRN+ [24]	-	3.31	96.11	-
MFNet [25]	3.43	3.74	97.98	-
TridentSE [26]	3.44	-	97.86	-
FRCRN_SE_16K	3.24	3.66	97.73	4.03
MossFormerGAN_SE_16K	**3.57**	**3.88**	**98.05**	**4.05**

Table 2: *Speech enhancement VoiceBank+DEMAND benchmark (48 kHz). The test set includes 824 clips from two speakers with multiple SNR levels (-7.5 dB, -2.5 dB, 2.5 dB, and 7.5 dB)*

Model	PESQ	NB_PESQ	SISDR	MCD	P808 MOS
Noisy	1.97	2.87	8.39	5.41	3.07
Resemble_Enhance	2.84	3.58	12.42	1.54	**3.53**
DeepFilterNet [8]	3.03	3.71	15.71	1.77	3.47
MossFormer2_SE_48K	**3.15**	**3.77**	**19.36**	**0.53**	**3.53**

[5] https://librivox.org/

[6] http://wham.whisper.ai/

[7] https://freesound.org/

[8] https://github.com/audiolabs/rir-generator

Table 3: *Speech separation benchmarks (8 kHz & 16 kHz), each containing 3000 utterances: LRS2-2Mix (noisy and reverberant, SNR -5 dB to 5 dB), WSJ0-2Mix (clean), Libri2Mix (noisy, LUFS -25 dB to -33 dB), and WHAM! (noisy, SNR -6 dB to 3 dB).*

Model	LRS2-2Mix	WSJ0-2Mix	Lirbi2Mix	WHAM!
DualPathRNN [32]	12.7	18.8	16.1	13.7
SepFormer [33]	13.5	20.4	17.0	14.4
TDANet [34]	14.2	18.5	17.4	15.2
TF-GridNet [35]	-	**22.8**	19.8	16.9
SPMamba [36]	-	22.5	**19.9**	**17.4**
MossFormer2_SS_16K	**15.5**	22.0	16.7	**17.4**

Table 4: *Speech super-resolution performance evaluation.*

Model	PESQ	16 kHz	24 kHz	32 kHz	48 kHz
Origin	1.97	2.80	2.60	2.29	1.46
Enhanced	**3.15**	**1.93**	**1.52**	**1.50**	**1.42**

training, we implemented architecture-specific model initialization and optimizer selection. Efficient data loading is achieved via distributed sampling, while gradient accumulation allows for larger effective batch sizes. Gradient clipping was adopted to enhance training stability and efficiency. We utilized a learning rate schedule with fine-tuning and halving and enforced regular check-pointing and logging for monitoring and resuming training.

The training script uses different loss functions tailored to each model. FRCRN_SE_16K combines SiSNR and the complex-valued masking MSE loss for spectral fidelity and temporal consistency [11]. Mossformer2_SE_48K uses the masking MSE loss [29], considering both magnitude and phase. MossformerGan_SE_16K employs an adversarial approach, combining generator loss (MSE on discriminator output), magnitude loss (MSE on spectrograms), real/imaginary loss (MSE on STFT coefficients), time domain loss, and discriminator loss incorporating PESQ for perceptual alignment [30]. These diverse loss functions guide model training toward optimal performance across multiple audio quality dimensions. MossFormer2_SS_16K uses Permutation Invariant Training (PIT) with SI-SNR loss [31]. The optimization for MossFormer2_SR_48K combines GAN loss corresponding to the three discriminator types of MSD, MPD, and MBD, multi-scale mel-spectrogram loss, and feature matching loss [14].

3.2. Model performance

We provide the evaluation performance for each model below.
SE: Our 16 kHz speech enhancement models are compared against DCCRN+ [24], MFNet [25], and TridentSE [26] on the DNS Challenge 2020 benchmark, while our 48 kHz model is benchmarked against Resemble_Enhance and DeepFilterNet on VoiceBank+DEMAND. Results are shown in Tables 1 and 2, demonstrating competitive performance despite training on general-purpose datasets.
SS: Speech separation performance was evaluated on LRS2-2Mix (16 kHz), WSJ0-2Mix (8 kHz), Libri2Mix (8 kHz), and WHAM! (8 kHz) benchmarks, comparing against DualPathRNN [32], SepFormer [33], TDANet [34], TF-GridNet [35], and SPMamba [36] models (results from published works). MossFormer2_SS_16K, evaluated on resampled 16 kHz audio from our unified model (without dataset-specific retraining), achieved competitive results on LRS2-2Mix, WSJ0-2Mix, and WHAM! as shown in Table 3. However, its lower performance on Libri2Mix may be due to the dataset's inherently lower speech quality.

Table 5: *Face-conditioned speaker extraction performance on VoxCeleb2 benchmark, results are reported in dB on 2 speakers mixtures (2-mix) and 3 speakers mixtures (3-mix).*

Model	Params	2-mix SI-SNRi	2-mix SNRi	3-mix SI-SNRi	3-mix SNRi
AV-ConvTasNet [37]	20.2	10.6	10.9	9.8	10.2
MuSE [38]	25.0	11.7	12.0	11.6	12.2
reentry [39]	18.8	12.6	12.9	12.6	13.1
AV-DPRNN [40]	**15.3**	11.5	11.8	10.5	11.0
AV-TFGridNet [16]	20.8	13.7	14.1	14.2	14.6
AV-MossFormer2	68.5	**14.6**	**14.9**	**15.5**	**16.0**

SR: We demonstrated the effectiveness of our speech super-resolution model, MossFormer2_SR_48K, using the VoiceBank+DEMAND 48 kHz test set. For super-resolution evaluation, the test set was downsampled to 16 kHz, 24 kHz, and 32 kHz. Recognizing that speech quality is impacted by both lower sampling rates and background noise, we also incorporated our speech enhancement model, MossFormer2_SE_48K, to reduce noise prior to super-resolution processing. Results are presented in Table 4. While MossFormer2_SE_48K significantly improved the 16 kHz PESQ score, its impact on LSD was marginal (see the last column). Consequently, the LSD improvements observed at 16 kHz, 24 kHz, and 32 kHz are primarily due to MossFormer2_SR_48K.
AVSE: To standardize audio-visual speaker extraction (AVSE) evaluation, we introduced the VoxCeleb2 mixture dataset, and established benchmarks shown in Table 5. AV-MossFormer2 outperforms the state-of-the-art AV-TFGridNet on both 2-speaker and 3-speaker mixtures. Benchmarks for gesture-conditioned (YGD) and brain-EEG-conditioned (KUL) speaker extraction are also provided in the repository.

4. Discussion and future plan

We presented ClearerVoice-Studio serving as a valuable bridge between advanced speech processing research and practical applications. Beyond the presented evaluations, ClearerVoice-Studio is available for live demos on HuggingFace and ModelScope, enabling users to experiment with real-world recordings. We have also generated a 48 kHz version of the LJSpeech-1.1 TTS corpus based on ClearerVoice-Studio, now publicly available[9]. ClearerVoice-Studio has rapidly gained traction within the research and development community, evidenced by over 2,800 GitHub stars and 200+ forks, signifying widespread interest and active adaptation.

ClearerVoice-Studio's future development will focus on integrating cutting-edge models and algorithms (e.g., diffusion models), expanding support for additional tasks and modalities (e.g., multi-channel audio and video), and enhancing training tools through streamlined pipelines, improved visualization, and automated hyperparameter tuning. The development will also prioritize real-time processing and edge deployment, while continuing to foster community engagement. These initiatives will ensure ClearerVoice-Studio remains a dynamic and influential resource within the speech processing community.

[9]https://huggingface.co/datasets/alibabasglab/LJSpeech-1.1-48kHz

5. References

[1] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and T. Sainath, "Deep learning for audio signal processing," *IEEE J. Sel. Top. Signal Process.*, vol. 13, no. 2, pp. 206–219, 2019.

[2] X. Zhang, Q. Zhang, H. Liu, T. Xiao, X. Qian, B. Ahmed, E. Ambikairajah, H. Li, and J. Epps, "Mamba in Speech: Towards an alternative to self-attention," *preprint arXiv:2405.12609*, 2025.

[3] Q. Zhang, A. Nicolson, M. Wang, K. K. Paliwal, and C. Wang, "DeepMMSE: A deep learning approach to mmse-based noise power spectral density estimation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 1404–1415, 2020.

[4] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021. [Online]. Available: https://arxiv.org/abs/2106.04624

[5] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-end speech processing toolkit," in *Proc. Interspeech*, 2018.

[6] H. Chen, S. Wu, C. Wang, J. Du, C.-H. Lee, S. M. Siniscalchi, S. Watanabe, J. Chen, O. Scharenborg, Z.-Q. Wang, B.-C. Yin, and J. Pan, "Summary on the multimodal information-based speech processing 2023 challenge," in *Proc. ICASSPW*, 2024, pp. 123–124.

[7] M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera, F.-R. Stöter, M. Hu, J. M. Martín-Doñas, D. Ditter, A. Frank, A. Deleforge, and E. Vincent, "Asteroid: the PyTorch-based audio source separation toolkit for researchers," in *Proc. Interspeech*, 2020.

[8] H. Schröter, A. N. Escalante, , T. Rosenkranz, and A. Maier, "DeepFilterNet: Perceptually motivated real-time speech enhancement," in *Proc. Interspeech*, 2023.

[9] H. Liu, K. Chen, Q. Tian, W. Wang, and M. D. Plumbley, "AudioSR: Versatile audio super-resolution at scale," in *Proc. ICASSP*, 2024, pp. 1076–1080.

[10] C. Jung, S. Lee, J.-H. Kim, and J. S. Chung, "FlowAVSE: Efficient audio-visual speech enhancement with conditional flow matching," in *Proc. Interspeech*, 2024.

[11] S. Zhao, B. Ma, K. N. Watcharasupat, and W.-S. Gan, "FRCRN: Boosting feature representation using frequency recurrence for monaural speech enhancement," in *Proc. ICASSP*, 2022.

[12] S. Zhao, Y. Ma, C. Ni, C. Zhang, H. Wang, T. H. Nguyen, K. Zhou, J. Yip, D. Ng, and B. Ma, "MossFormer2: Combining transformer and rnn-free recurrent network for enhanced time-domain monaural speech separation," in *Proc. ICASSP*, 2024.

[13] S. Zhao and B. Ma, "MossFormer: Pushing the performance limit of monaural speech separation using gated single-head transformer with convolution-augmented joint self-attentions," in *Proc. ICASSP*, 2023.

[14] S. Zhao, K. Zhou, Z. Pan, Y. Ma, C. Zhang, and B. Ma, "Hifi-SR: A unified generative transformer-convolutional adversarial network for high-fidelity speech super-resolution," in *Proc. ICASSP*, 2025.

[15] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. NeurIPS*, 2020.

[16] 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.

[17] Z. Pan, X. Qian, and H. Li, "Speaker extraction with co-speech gestures cue," *IEEE Signal Process. Lett.*, vol. 29, pp. 1467–1471, 2022.

[18] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "SpEx+: A complete time domain speaker extraction network," in *Proc. Interspeech*, 2020, pp. 1406–1410.

[19] Z. Pan, M. Borsdorf, S. Cai, T. Schultz, and H. Li, "NeuroHeed: Neuro-steered speaker extraction using EEG signals," *arXiv:2307.14303*, 2023.

[20] C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, and S. B. *et al.*, "The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," in *Proc. Interspeech*, 2020.

[21] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK Corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:213060286

[22] J. F. Gemmeke, D. P. W. 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.

[23] J. Thiemann, N. Ito, and E. Vincent, "DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments," Jun. 2013. [Online]. Available: https://doi.org/10.5281/zenodo.1227121

[24] S. Lv, Y. Hu, S. Zhang, and L. Xie, "DCCRN+: Channel-wise subband dccrn with snr estimation for speech enhancement," in *Proc. Interspeech*, 2021.

[25] L. Liu, H. Guan, J. Ma, W. Dai, G. Wang, and S. Ding, "A mask free neural network for monaural speech enhancement," in *Proc. Interspeech*, 2023.

[26] D. Yin, Z. Zhao, C. Tang, Z. Xiong, and C. Luo, "TridentSE: Guiding speech enhancement with 32 global tokens," in *Proc. Interspeech*, 2023.

[27] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A corpus derived from librispeech for text-to-speech," in *Proc. Interspeech*, 2019.

[28] H. Larochelle, Y. Bengio, J. Louradour, and P. Lamblin, "Exploring strategies for training deep neural networks," *Journal of Machine Learning Research*, vol. 10, no. 1, pp. 1–40, 2009.

[29] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. ICASSP*, 2015, pp. 708–712.

[30] R. Cao, S. Abdulatif, and B. Yang, "CMGAN: Conformer-based metric gan for speech enhancement," in *Proc. Interspeech*, 2022.

[31] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. ICASSP*, 2017.

[32] 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.

[33] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *Proc. ICASSP*, 2021.

[34] K. Li, R. Yang, and X. Hu, "An efficient encoder-decoder architecture with top-down attention for speech separation," in *Proc. ICLR*, 2023.

[35] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "TF-GridNet: Making time-frequency domain models great again for monaural speaker separation," in *Proc. ICASSP*, 2023.

[36] K. Li, G. Chen, R. Yang, and X. Hu, "SPMamba: State-space model is all you need in speech separation," *arXiv:2404.02063*, 2024.

[37] J. Wu, Y. Xu, S. Zhang, L. Chen, M. Yu, L. Xie, and D. Yu, "Time domain audio visual speech separation," in *Proc. ASRU*, 2019.

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

[39] ——, "Selective listening by synchronizing speech with lips," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 1650–1664, 2022.

[40] Z. Pan, M. Ge, and H. Li, "USEV: Universal speaker extraction with visual cue," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 3032–3045, 2022.