

ACTIVE SPEECH ENHANCEMENT: BEYOND PASSIVE DENOISING DECLIPPING AND DEREVERBERATION

Anonymous authors

Paper under double-blind review

ABSTRACT

We introduce a new paradigm for active sound modification: *Active Speech Enhancement* (ASE). While Active Noise Cancellation (ANC) algorithms focus on suppressing external interference and traditional speech enhancement passively reconstructs degraded speech signals, ASE goes further by actively shaping the speech signal, both attenuating unwanted noise components and amplifying speech-relevant frequencies to improve intelligibility and perceptual quality. To enable this, we propose a novel Transformer-Mamba-based architecture, along with a task-specific loss function designed to jointly optimize interference suppression and signal enrichment in an acoustic environment. Our method outperforms existing baselines across multiple speech processing tasks, including denoising, dereverberation, and declipping, demonstrating the effectiveness of active, targeted modulation in challenging acoustic environments. A demo page and source code are provided in the Supplementary Materials.

1 INTRODUCTION

Traditional speech enhancement and noise control are fundamental audio processing tasks. Traditional speech enhancement aims to passively improve the perceptual quality and intelligibility of speech signals by mitigating degradations such as background noise, distortion, clipping, and reverberation. Classic approaches—including spectral subtraction, Wiener filtering, and statistical model-based methods—have achieved varying degrees of success but often falter in highly non-stationary noise environments (Boll, 2003; Lim & Oppenheim, 1978; Paliwal et al., 2012). Recent advances in deep learning have, however, yielded state-of-the-art performance: convolutional neural networks (CNNs) (Pascual et al., 2017; Rethage et al., 2018; Pandey & Wang, 2018), recurrent neural networks (RNNs) (Hu et al., 2020), generative adversarial networks (GANs) (Fu et al., 2019; 2021; Kim et al., 2021; Shin et al., 2023; Shetu et al., 2025), Transformers (Wang et al., 2021; de Oliveira et al., 2022; Zhang et al., 2022b; Cao et al., 2022; Ye & Wan, 2023; Zhang et al., 2024), and diffusion models (Guimarães et al., 2025; Lu et al., 2022; Welker et al., 2022; Richter et al., 2023; Lemerrier et al., 2023; Tai et al., 2023; Ayilo et al., 2024) demonstrate exceptional results on benchmarks for denoising, dereverberation, and declipping.

Active noise cancellation takes a complementary approach by generating an anti-noise signal to interfere with unwanted noise destructively. Pioneering work dating back to Lueg’s first patent in 1936 introduced the concept of adaptive feedforward ANC (Lueg, 1936), which was later refined through advances in adaptive filtering (e.g., LMS, FxLMS), multi-channel algorithms, and applications in headphones and enclosure systems (Nelson & Elliott, 1991; Fuller et al., 1996; Hansen et al., 1997; Kuo & Morgan, 1999; Zhang & Wang, 2021; Park et al., 2023; Mostafavi & Cha, 2023; Cha et al., 2023; Singh et al., 2024; Pike & Cheer, 2023; Mishaly et al., 2025). While ANC excels at suppressing predictable or narrowband noise, it does not actively modify the speech content itself.

We propose a new paradigm—*Active Speech Enhancement* (ASE) that unifies the goals of traditional speech enhancement and active noise control. Unlike conventional ANC, which solely targets noise suppression, ASE actively shapes the speech signal by simultaneously attenuating interfering components and amplifying speech-related frequencies. This dual-action approach not only reduces noise but also emphasizes speech and improves perceptual quality under challenging acoustic conditions. We make four key contributions. First, we formalize the ASE task and describe appropriate evaluation metrics that capture both noise suppression and speech enhancement. Second, we

introduce a Transformer-Mamba-based model that generates an active modification signal, leveraging self-attention to capture long-range dependencies in time–frequency representations. Third, we design a joint suppression–enrichment loss that balances interference removal and signal enrichment, combining spectral, perceptual, and adversarial objectives to drive optimal ASE performance. Fourth, we conduct a comprehensive evaluation, showing that our method outperforms adapted baselines across multiple ASE tasks—including denoising, dereverberation, and declipping—with significant gains in metrics such as PESQ (Rix et al., 2001).

2 RELATED WORK

2.1 TRADITIONAL SPEECH ENHANCEMENT

Recent advances in deep learning have yielded substantial improvements in traditional speech enhancement. Pandey and Wang (Pandey & Wang, 2018) proposed a CNN-based autoencoder that applies convolutions directly to the raw waveform while computing the loss in the frequency domain. SEGAN (Pascual et al., 2017) employs strided convolutional layers in a generative adversarial framework. Rethage et al. (2018) developed a WaveNet-inspired model that predicts multiple waveform samples per step to reduce computational cost. Recurrent architectures have also been explored. Hu et al. (2020) presented DCCRN, which integrates complex-valued convolutional and recurrent layers to process spectrogram inputs. A real-time causal model based on an encoder–decoder with skip connections was proposed by Defossez et al. (2020), operating directly on the raw waveform and optimized in both time and frequency domains.

MetricGAN (Fu et al., 2019) and its successor MetricGAN+ (Fu et al., 2021) incorporate evaluation metrics such as PESQ (Perceptual Evaluation of Speech Quality) (Rix et al., 2001) and STOI (Short-Time Objective Intelligibility) (Kim et al., 2021) into the adversarial loss. Kim et al. (2021) further enhance this approach by introducing a multiscale discriminator operating at different sampling rates alongside a generator that processes speech at multiple resolutions.

Transformer-based models have recently gained prominence. Wang et al. (2021) proposed a two-stage transformer network (TSTNN) that outperforms earlier time- and frequency-domain methods. CMGAN (Cao et al., 2022) adapts the Conformer backbone (Gulati et al., 2020) for enhancement, and de Oliveira et al. (2022) replace the learned encoder of SepFormer (Subakan et al., 2021) with long-frame STFT (Short-Time Fourier Transform) inputs, reducing sequence length and lowering computational cost without compromising perceptual quality.

More recently, diffusion-based approaches have emerged as a powerful generative paradigm. Lu et al. (2022) introduced a conditional diffusion probabilistic model that learns a parameterized reverse diffusion process conditioned on the noisy input. Welker et al. (2022) extended score-based models to the complex STFT domain, learning the gradient of the log-density of clean speech coefficients. Richter et al. (2023) formulate enhancement as a stochastic differential equation, initializing reverse diffusion from a mixture of noisy speech and Gaussian noise and achieving high-quality reconstructions in only 30 steps. Lemerrier et al. (2023) propose a *stochastic regeneration* method that leverages an initial predictive-model estimate to guide a reduced-step diffusion process, mitigating artifacts and reducing computational cost by an order of magnitude while maintaining quality.

2.2 ACTIVE AUDIO CANCELLATION

Recently, deep learning approaches have demonstrated remarkable results in ANC algorithms. Zhang & Wang (2021) introduced DeepANC, which employs a convolutional long short-term memory (Conv-LSTM) network to jointly estimate amplitude and phase responses from microphone signals. Subsequently, attention-driven ANC frameworks integrating attentive recurrent networks were proposed to enable real-time adaptation and low-latency operation (Zhang et al., 2022a).

A selective fixed-filter ANC (SFANC) framework was developed to leverage a two-dimensional CNN for optimal control-filter selection on a mobile co-processor and a lightweight one-dimensional CNN for time-domain noise classification, yielding superior attenuation of real-world non-stationary headphone noise (Shi et al., 2022). Luo et al. (2022) proposed a hybrid SFANC–FxNLMS that first applies a similar approach as SFANC for each noise frame and then applies the FxNLMS algorithm for real-time coefficient adaptation, thereby combining the rapid response of SFANC with the low

steady-state error and robustness of adaptive optimization. Heuristic algorithms—such as bee colony optimization (Ren & Zhang, 2022) and genetic algorithms (Zhou et al., 2023)—have been explored to avoid gradient-based learning.

Other studies have applied recurrent convolutional networks (Park et al., 2023; Mostafavi & Cha, 2023; Cha et al., 2023) and fully connected neural networks (Pike & Cheer, 2023) to ANC. Autoencoder-based encodings have been used to extract latent features for improved robustness (Singh et al., 2024). Efforts in SFANC have extended to synthesizing optimized filter banks via unsupervised methods (Luo et al., 2024), while advancements in multichannel setups continue to leverage spatial diversity through deep controllers (Shi et al., 2024). Multichannel configurations have been further enhanced by refined deep controllers that learn inter-channel relationships for improved noise attenuation (Zhang & Wang, 2023; Antoñanzas et al., 2023; Xiao et al., 2023; Zhang et al., 2023b; Shi et al., 2023), and attention-driven frameworks have been investigated for low-latency operation (Zhang et al., 2023a).

3 BACKGROUND

We first examine a feedforward ANC algorithm that employs a single error microphone to lay the foundation for our new ASE framework. In the ANC framework (Figure 1a), the *primary path*, characterized by the transfer function $P(z)$, models the acoustic propagation from the disturbance source to the error microphone. The *secondary path*, denoted by $S(z)$, describes the transfer from the loudspeaker to the error microphone. Let $x(n)$ denote the reference signal applied to the ANC system. The primary signal $d(n)$ is obtained by filtering $x(n)$ through the primary path:

$$d(n) = P(z) * x(n), \quad (1)$$

where $*$ denotes the convolution operation. The error microphone captures the residual signal $e(n)$, representing the difference between the original disturbance and the cancellation signal. Both $x(n)$ and $e(n)$ are used by the ANC algorithm to compute the canceling signal $y(n)$. The loudspeaker implements $y(n)$ according to its electro-acoustic transfer function $f_{LS}\{\cdot\}$. After propagation through the secondary path, the anti-signal (or cancellation signal) is given by

$$a(n) = S(z) * f_{LS}(y(n)). \quad (2)$$

The error signal is defined formally as the difference between the primary signal and the anti-signal:

$$e(n) = d(n) - a(n). \quad (3)$$

The objective of the ANC algorithm is to generate $y(n)$ such that $e(n)$ is minimized, ideally achieving $e(n) = 0$, which corresponds to complete cancellation of the disturbance. In contrast, the **ASE framework** uses the primary and anti-signals to construct an enhanced signal:

$$eh(n) = d(n) + a(n). \quad (4)$$

While ANC aims to eliminate the disturbance, ASE seeks to recover clean speech from a noisy mixture of distorted speech $x(n)$. The objective of the ASE task is to generate $eh(n)$ such that its deviation from the clean target signal $c(n)$, i.e., $eh(n) - c(n)$, is minimized. As illustrated in Figure 1b, the feedforward ASE setup comprises a disturbance source, a reference signal path, and a control filter operating through the secondary path. Given the nature of the task, the error microphone serves as the modification microphone in our framework.

Our work adapts three speech distortion types, previously defined by VoiceFixer (Liu et al., 2022) for general speech restoration, to the context of our ASE framework. Specifically, our ASE-TM model targets the restoration of speech $s(n)$ degraded by: **(i) Additive noise**: This common distortion, where unwanted background sounds obscure the speech, is modeled as the sum of the clean speech signal $s(n)$ and a noise signal $n(n)$:

$$d_{\text{noise}}(s(n)) = s(n) + n(n). \quad (5)$$

(ii) Reverberation: Caused by sound reflections in an enclosure, reverberation blurs speech signals. It is modeled by convolving the speech signal $s(n)$ with a room impulse response (RIR) $r(n)$:

$$d_{\text{rev}}(s(n)) = s(n) * r(n). \quad (6)$$

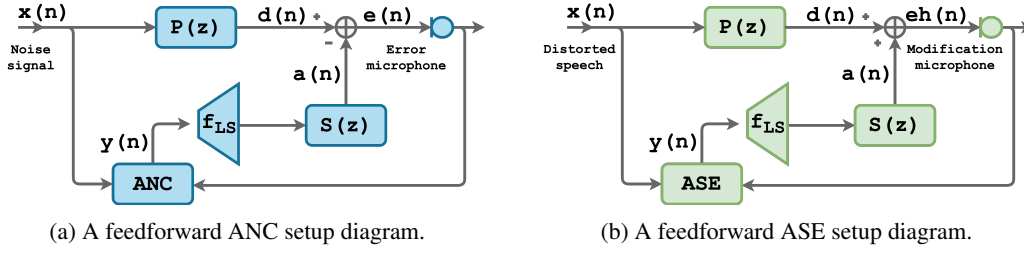


Figure 1: Comparison of feedforward ANC and ASE setups.

(iii) Clipping: This distortion arises when signal amplitudes exceed the maximum recordable level, typically due to microphone limitations. Clipping truncates the signal $s(n)$ within a certain range $[-\eta, +\eta]$:

$$d_{\text{clip}}(s(n)) = \max(\min(s(n), \eta), -\eta), \quad \eta \in [0, 1]. \quad (7)$$

This leads to harmonic distortions and can degrade speech intelligibility.

To assess the performance of ASE-TM across these enhancement tasks, we employ a suite of established objective metrics. Consistent with the evaluation protocol in SEMamba (Chao et al., 2024), these include the Wide-band PESQ (Rix et al., 2001), STOI (Taal et al., 2010), and the composite measures CSIG (predicting signal distortion), CBAK (predicting background intrusiveness), and COVL (predicting overall speech quality) (Hu & Loizou, 2007). Furthermore, we incorporate the Normalized Mean Square Error (NMSE), a traditionally well-established metric in the ANC task. The NMSE between a target signal $u(n)$ and an estimated signal $v(n)$ is defined in decibels (dB) as:

$$\text{NMSE}[u, v] = 10 \cdot \log_{10} \left(\frac{\sum_{n=1}^M (u(n) - v(n))^2}{\sum_{n=1}^M u(n)^2} \right), \quad (8)$$

where M is the total number of samples. In our evaluations, $u(n)$ represents the clean target speech $c(n)$, and $v(n)$ is the enhanced speech $eh(n)$ produced by our model (the precise definition of $c(n)$ for each task is detailed in Section 4.1).

4 METHOD

4.1 ASE-TM ARCHITECTURE OVERVIEW

The proposed model, ASE Transformer-Mamba (ASE-TM, Figure 2), adopts and extends the fundamental structure of the SEMamba architecture (Chao et al., 2024), which consists of a dense encoder, a series of Time-Frequency Mamba (TFMamba) blocks (Xiao & Das, 2024), and parallel magnitude and phase decoders. A notable distinction of our ASE-TM model is the utilization of Mamba2 blocks (Dao & Gu, 2024) within these TFMamba pathways, in contrast to the original SEMamba architecture, which employed an earlier version of Mamba (Gu & Dao, 2023). This choice is motivated by the potential improvements in efficiency offered by Mamba2.

The input noisy waveform, $x(n)$, sampled at a rate of F_s , is processed through an STFT. The STFT employs a Hann window of N_{win} samples, a hop length of N_{hop} samples, and an N_{FFT} -point FFT, resulting in $N_{\text{freq}} = \lfloor N_{\text{win}}/2 \rfloor + 1$ frequency features per frame. Its magnitude and phase components are horizontally stacked and fed into the network. The dense encoder utilizes convolutional layers and dense blocks to extract initial features from the stacked magnitude and phase, outputting a representation with C_{enc} channels, each with N_{enc} features.

The core of the temporal and spectral modeling is based on N_{tf} TFMamba blocks. Each TFMamba block contains separate Mamba-based pathways (time-mamba and freq-mamba) employing bidirectional Mamba layers to capture dependencies across time and frequency dimensions, respectively (Chao et al., 2024; Xiao & Das, 2024).

Following the initial $N_{\text{tf}}/2$ TFMamba blocks, inspired by hybrid approaches like Jamba (Lieber et al., 2024), we introduce an attention-based block. Before applying attention, the feature representation, with C_{enc} channels, undergoes dimensionality reduction. A 2D convolution with a kernel size

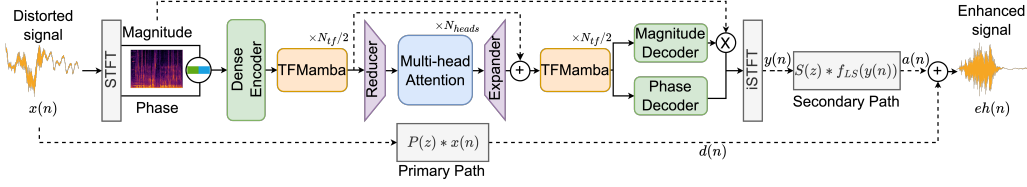


Figure 2: ASE-TM Architecture.

of $(1, N_{\text{enc}}/2 + 1)$ reduces the channel dimension from C_{enc} to $C_{\text{enc}}/4$ and each channel features size from N_{enc} to $N_{\text{enc}}/2$. This results in a compact representation of size $N_{\text{enc}}/2 \times C_{\text{enc}}/4$ for the attention layer. In addition, positional encoding is applied to this compact representation. A standard Multi-Head Attention layer with N_{heads} heads is then used on this reduced representation to weigh features based on global context. Following the attention layer, an expansion module employing a transposed 2D convolution, also with a kernel size of $(1, N_{\text{enc}}/2 + 1)$, is used to restore the channel dimension to C_{enc} and expand the feature dimension back towards N_{enc} before passing the features to the remaining $N_{\text{tf}}/2$ TFMamba blocks. An additional step before applying the remaining TFMamba blocks is the use of a residual connection that sums the feature representations from before the dimensionality reduction with those after the attention and expansion modules.

The magnitude and phase decoders retain the structure used in SEMamba, employing dense blocks and convolutional layers (including transposed convolutions) to reconstruct the target representation—not before applying a residual connection that performs element-wise multiplication between the original STFT magnitude and the predicted magnitude. However, instead of predicting the enhanced spectra directly, the network is trained to output the complex spectrum of the cancelling signal $y(n)$. This signal $y(n)$, after undergoing the electro-acoustic transfer function $f_{\text{LS}}\{\cdot\}$ and propagation through the secondary path $S(z)$, becomes the anti-signal $a(n)$ (as defined in Eq. 2). This anti-signal $a(n)$ is then summed with the primary path signal $d(n)$ to produce the final enhanced signal $eh(n)$ (as defined in Eq. 3).

4.2 OPTIMIZATION OBJECTIVE

The primary goal of the ASE-TM model is to generate an enhanced signal $eh(n)$ that is as close as possible to a clean target speech signal $c(n)$. The definition of this target $c(n)$ varies based on the specific enhancement task. For **additive noise reduction**, $c(n)$ is the clean speech signal convolved with the primary path $P(z)$, representing the clean signal as perceived at the modification microphone. For **dereverberation** and **declipping**, $c(n)$ is the original anechoic, unclipped clean speech signal, prior to any acoustic path effects or clipping distortion.

The training of ASE-TM largely follows the multi-level loss framework established in SEMamba and originating from MP-SENet (Lu et al., 2023). This framework combines several loss components. Our approach incorporates this established framework with specific modifications. The overall generator loss $\mathcal{L}_G$ is a weighted sum of the following components:

1. **Time-Domain Loss ($\mathcal{L}_T$):** We employ a combination of L1 and L2 distances between the enhanced waveform $eh(n)$ and the target waveform $c(n)$:

$$\mathcal{L}_T = \|eh(n) - c(n)\|_1 + \|eh(n) - c(n)\|_2^2. \quad (9)$$

This hybrid loss aims to leverage the robustness of L1 to outliers and the smoothness encouraged by L2.

2. **Magnitude Spectrum Loss ($\mathcal{L}_{\text{Mag}}$):** Similar to the time-domain loss, a combined L1 and L2 loss on the magnitude spectra is applied. If EN_m and C_m are the magnitude spectra of $eh(n)$ and $c(n)$ respectively, then:

$$\mathcal{L}_{\text{Mag}} = \|EN_m - C_m\|_1 + \|EN_m - C_m\|_2^2. \quad (10)$$

This contrasts with the L2 loss typically used in MP-SENet for this component.

3. **Complex Spectrum Loss ($\mathcal{L}_{\text{Com}}$):** This loss penalizes differences in the STFT domain. It is the sum of L2 losses on the real and imaginary parts of the STFTs of $eh(n)$ and $c(n)$.

4. **Anti-Wrapping Phase Loss** ($\mathcal{L}_{\text{Pha}}$): Includes instantaneous phase loss, group delay loss, and instantaneous angular frequency loss to optimize the phase spectrum directly, addressing phase wrapping issues.
5. **Metric-Based Adversarial Loss** ($\mathcal{L}_{\text{Met}}$): A discriminator trained to predict a perceptual metric (e.g., PESQ), guiding the generator to produce outputs that score well on it.
6. **Consistency Loss** ($\mathcal{L}_{\text{Con}}$): We incorporate a consistency loss. This loss minimizes the discrepancy between the complex spectrum directly output by the model’s decoders (magnitude and phase) and the complex spectrum obtained by applying STFT to the time-domain waveform $eh(n)$ that results from an inverse STFT of the initially predicted spectrum.

The total generator loss is then defined with the hyperparameter γ as follows:

$$\mathcal{L}_G = \gamma_1 \mathcal{L}_T + \gamma_2 \mathcal{L}_{\text{Mag}} + \gamma_3 \mathcal{L}_{\text{Com}} + \gamma_4 \mathcal{L}_{\text{Met}} + \gamma_5 \mathcal{L}_{\text{Pha}} + \gamma_6 \mathcal{L}_{\text{Con}}. \quad (11)$$

5 EXPERIMENTS

5.1 DATASETS AND TASK GENERATION

We conduct evaluations across three primary speech restoration tasks: additive noise reduction, dereverberation, and declipping. For **additive noise reduction**, we use the VoiceBank-DEMAND dataset (Botinhao et al., 2016), a standard benchmark in speech enhancement. This dataset combines clean speech from the VoiceBank corpus (Veaux et al., 2013) with various non-stationary noises from the DEMAND database (Thiemann et al., 2013). Our training set consists of utterances from 28 speakers with 10 different noise types at Signal-to-Noise Ratios (SNRs) of 0, 5, 10, and 15 dB. We used two speakers from the training set as the validation set. The test set comprises 824 utterances from 2 unseen speakers, mixed with 5 unseen noise types at SNRs of 2.5, 7.5, 12.5, and 17.5 dB.

The datasets for the **dereverberation** and **declipping** tasks are generated using the clean speech utterances of the speakers available in the VoiceBank corpus. **Dereverberation**: Reverberant speech is synthesized by convolving the clean VoiceBank utterances with Room Impulse Responses (RIRs) as defined in Eq. 6 using the SpeechBrain package (Ravanelli et al., 2024). For training, we randomly sample RIRs from the training portion of the RIR dataset provided alongside VoiceFixer (Liu et al., 2022). For the test set, a fixed and distinct set of RIRs (from the VoiceFixer RIR test set) is applied to the clean test utterances to ensure consistent evaluation conditions. **Declipping**: Clipped speech signals are generated by applying a clipping threshold η to the clean utterances according to Eq. 7. During training, the clipping ratio η is uniformly sampled from the range $[0.1, 0.5]$ for each utterance to expose the model to varying degrees of distortion. For testing, a specific clipping threshold is used.

5.2 ACOUSTIC PATH SIMULATION

To emulate the acoustic environment for the ASE framework, we simulate the primary path $P(z)$ and secondary path $S(z)$. Our simulation setup is based on previous setups for the ANC task (Zhang & Wang, 2021; Zhang et al., 2023a), modeling a rectangular enclosure with dimensions of $3 \times 4 \times 2$ meters (width $\times$ length $\times$ height). Room Impulse Responses (RIRs) are generated using the image method (Allen & Berkley, 1979), implemented with a Python-based RIR generator package (Habets, 2006) with high-pass filtering. The modification microphone, capturing $eh(n)$, is at $[1.5, 3, 1]$ meters. The reference microphone, capturing $x(n)$, is at $[1.5, 1, 1]$ meters, and the cancellation loudspeaker, which outputs the signal leading to $a(n)$, is at $[1.5, 2.5, 1]$ meters within the enclosure. The RIR length for both $P(z)$ and $S(z)$ is $L_{\text{RIR}} = 512$ taps. The non-linear characteristics of the loudspeaker are modeled using the Scaled Error Function (SEF), defined as $f_{\text{LS}}\{y\} = \int_0^y \exp(-z^2/(2\lambda^2))dz$. Here, y is the loudspeaker input, and λ^2 controls the severity of the saturation non-linearity. Different λ^2 values simulate varying degrees of distortion, with larger values approaching linear behavior. To introduce variability during training, the room’s reverberation time (T_{60}) and λ^2 are randomly sampled from the sets $\{0.15, 0.175, 0.2, 0.225, 0.25\}$ seconds and $\{0.1, 1, 10, \infty\}$, respectively, for each training sample. For testing, fixed T_{60} and λ^2 are used.

Table 1: Average denoising results on the VoiceBank-DEMAND test set ($T_{60} = 0.25s$ and $\lambda^2 = \infty$).

Method	PESQ ($\uparrow$)	CSIG ($\uparrow$)	CBAK ($\uparrow$)	COVL ($\uparrow$)	STOI ($\uparrow$)	NMSE ($\downarrow$)
Noisy-speech	1.97	3.50	2.55	2.75	0.92	-8.44
THF-FxLMS	2.37	3.66	2.84	3.00	0.97	-15.32
DeepANC	1.48	1.99	2.19	1.69	0.93	-12.80
ARN	2.45	3.64	3.13	3.03	0.97	-20.64
ASE-TM	2.98	4.21	3.49	3.62	0.99	-21.76

5.3 MODEL HYPERPARAMETERS AND TRAINING

The ASE-TM model processes audio at $F_s = 16$ kHz. For the STFT, we use a Hann window of $N_{\text{win}} = 400$ samples, hop length of $N_{\text{hop}} = 100$ samples, and an $N_{\text{FFT}} = 400$ -point FFT. The dense encoder outputs a feature representation with $C_{\text{enc}} = 128$ channels, where each channel has a feature dimension of $N_{\text{enc}} = 100$. Our model employs a total of $N_{\text{tf}} = 8$ TFMamba blocks. The Multi-Head Attention layer within the attention-based block uses $N_{\text{heads}} = 10$ heads. Other internal architectural details for the TFMamba blocks and dense convolutional blocks largely follow the configurations presented in SEmamba. The ASE-TM model is trained for 350 epochs using the AdamW optimizer (Loshchilov & Hutter, 2017) with $\beta_1 = 0.8$ and $\beta_2 = 0.99$. The initial learning rate is set to 5×10^{-4} . We use a batch size of 4. Audio segments of 32,000 samples (equivalent to 2 seconds at 16 kHz) are used for training. The model parameters yielding the best performance on the validation set, evaluated based on the PESQ score, are saved for final testing. We used an NVIDIA RTX A6000 GPU (internal cluster). The training runtime of the ASE-TM model was ~ 10 days.

5.4 BASELINE METHODS

We compare ASE-TM with several established baseline methods commonly used in ANC. These include THF-FxLMS (Ghasemi et al., 2016), which is an extension to the traditional FxLMS algorithm (Kuo & Morgan, 1999), DeepANC that utilizes a convolutional LSTMs (Zhang & Wang, 2021), and ARN that incorporates an attention mechanism (Zhang et al., 2023a). These baseline methods were adapted and retrained or configured to the ASE framework across all tested tasks.

6 RESULTS AND ANALYSIS

6.1 ACTIVE DENOISING PERFORMANCE

The speech denoising performance on the VoiceBank-DEMAND dataset is in Table 1. Our ASE-TM model demonstrates superior performance, achieving a PESQ score of 2.98, significantly surpassing the baselines: THF-FxLMS achieved a PESQ of 2.37, and the deep learning-based ANC methods, DeepANC and ARN, yielded PESQ scores of 1.48 and 2.45, respectively. These results demonstrate a substantial gap between conventional ANC approaches and our ASE-TM, which benefits from actively shaping the speech signal in addition to noise suppression, as also evidenced by its leading scores in CSIG, CBAK, COVL, STOI, and a significantly better NMSE of -21.76 dB.

6.2 DEREVERBERATION AND DECLIPPING PERFORMANCE

ASE-TM’s efficacy was further evaluated on dereverberation and declipping tasks, with results presented in Tables 2 and 3, respectively. For dereverberation (Table 2), ASE-TM achieved a PESQ score of 2.43, a considerable improvement from the reverberant speech baseline (PESQ 1.60). In contrast, the adapted ANC baselines struggled; THF-FxLMS scored a PESQ of 1.43, while DeepANC and ARN achieved 1.06 and 1.35, respectively. This suggests that these methods, even when retrained or configured for the task, struggled to effectively adjust their processes to mitigate reverberation in the ASE framework. Similarly, in the declipping task, with a clipping threshold of $\eta = 0.25$ (Table 3), ASE-TM restored speech to a PESQ of 3.09 from an initial score of 2.17. The baseline methods again showed limited effectiveness: THF-FxLMS (PESQ 1.92), DeepANC (PESQ 1.05), and ARN (PESQ 1.67). These tasks, particularly where the target $c(n)$ is the original clean speech before any primary path effects, highlight the challenge and efficacy of the ASE approach in not just cancelling an interfering signal but actively restoring a desired signal characteristic.

Table 2: Average dereverberation results on the reverbed test set ($T_{60} = 0.25s$ and $\lambda^2 = \infty$).

Method	PESQ ($\uparrow$)	CSIG ($\uparrow$)	CBAK ($\uparrow$)	COVL ($\uparrow$)	STOI ($\uparrow$)	NMSE ($\downarrow$)
Reverbed-speech	1.60	2.60	1.88	2.02	0.80	2.00
THF-FxLMS	1.43	2.55	1.64	1.89	0.78	4.77
DeepANC	1.06	1.00	1.00	1.00	0.53	21.19
ARN	1.35	1.25	1.54	1.18	0.76	4.58
ASE-TM	2.43	3.71	2.67	3.07	0.93	-0.04

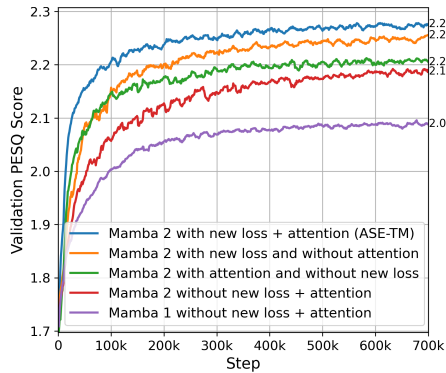
Table 3: Average declipping results on the clipped test set ($\eta = 0.25$, $T_{60} = 0.25s$, and $\lambda^2 = \infty$).

Method	PESQ ($\uparrow$)	CSIG ($\uparrow$)	CBAK ($\uparrow$)	COVL ($\uparrow$)	STOI ($\uparrow$)	NMSE ($\downarrow$)
Clipped-speech	2.17	3.49	2.51	2.82	0.89	-0.23
THF-FxLMS	1.92	3.35	2.36	2.62	0.88	-0.02
DeepANC	1.05	1.00	1.00	1.00	0.53	11.10
ARN	1.67	1.60	2.12	1.57	0.87	-0.31
ASE-TM	3.09	4.20	3.06	3.67	0.93	-1.70

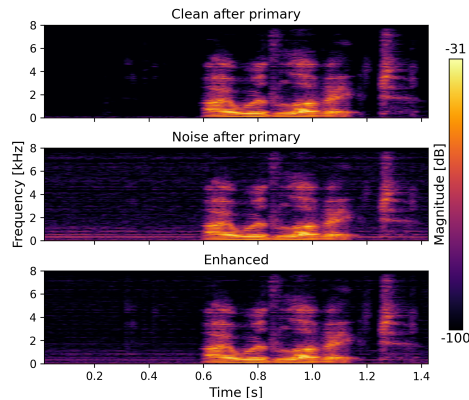
6.3 DENOISING ASE-TM MODEL ANALYSIS

An ablation study, presented in Figure 3a, investigates the contributions of our proposed loss function modifications, the attention mechanism, and the use of Mamba2 over Mamba1 to the ASE-TM model for the denoising task. The full model consistently achieves the highest validation PESQ score throughout training. Replacing Mamba1 with Mamba2 and the modified loss yielded the most considerable performance improvement among all evaluated components. Notably, configurations incorporating the attention mechanism demonstrate a faster convergence to higher performance levels, suggesting that attention aids in efficiently learning relevant features. Spectrogram analysis of a representative denoising example (Figure 3b) visually confirms the model’s effectiveness; the spectrogram of the enhanced signal closely mirrors that of the clean speech (after primary path), indicating successful noise suppression while preserving essential speech characteristics.

To assess robustness, ASE-TM was evaluated under varying acoustic conditions for the denoising task, with results in Table 4. This analysis focused on ASE-TM due to its significantly better performance over baselines in Table 1. The model shows consistent high performance across different T_{60} values under linear loudspeaker conditions ($\lambda^2 = \infty$), achieving a PESQ of 3.02 for $T_{60} = 0.15s$ and 3.13 for $T_{60} = 0.20s$. When strong non-linearities are introduced (e.g., $\lambda^2 = 0.1$ at $T_{60} = 0.25s$), the PESQ score is 2.74, still indicating robust performance. As λ^2 increases (less non-linearity), performance improves, reaching a PESQ of 2.97 for $\lambda^2 = 10$ at $T_{60} = 0.25s$.



(a) An ablation study. A moving average with window size = 10 was applied.



(b) An enhanced (denoised) signal spectrogram.

Figure 3: Model analysis of ASE-TM model for the denoising task.

Table 4: Average performance of ASE-TM (denoising task) under varying conditions (T_{60} and loudspeaker non-linearity factor λ^2) on the VoiceBank-DEMAND test set.

T_{60} (s)	λ^2	PESQ ($\uparrow$)	CSIG ($\uparrow$)	CBAK ($\uparrow$)	COVL ($\uparrow$)	STOI ($\uparrow$)	NMSE ($\downarrow$)
0.25	0.1	2.74	4.01	3.29	3.39	0.98	-20.21
0.25	1.0	2.92	4.17	3.44	3.57	0.99	-21.92
0.25	10	2.97	4.21	3.48	3.62	0.99	-22.37
0.15	∞	3.02	4.22	3.50	3.65	0.98	-22.31
0.20	∞	3.13	4.33	3.60	3.77	0.99	-22.88

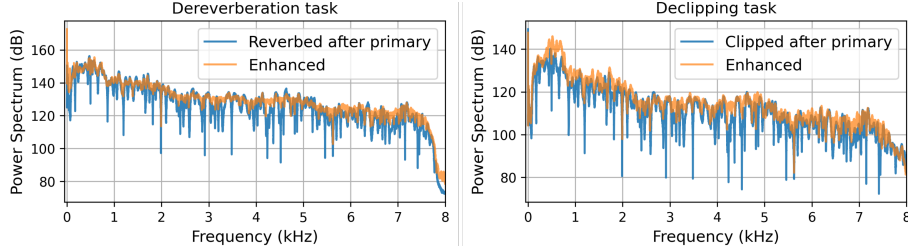


Figure 4: Power spectra for the dereverberation and declipping tasks over the entire test set.

6.4 RUNTIME ANALYSIS

To satisfy real-time constraints in active systems, we evaluated ASE-TM under a future-frame prediction strategy, following prior work (Zhang & Wang, 2021; Zhang et al., 2023a). In our setup, the causality condition $T_{\text{ASE-TM}} < T_p - T_s$ evaluates to $T_{\text{ASE-TM}} < \frac{2}{343} - \frac{0.5}{343} \approx 0.0043$ seconds, where T_p and T_s denote the acoustic delays of the primary and secondary paths, respectively. To accommodate the model’s inference latency, we predict 500 future frames (0.03125 seconds), remaining within real-time limits of our computational environment. Despite this future context, performance degradation is minimal: on the VoiceBank-DEMAND test set ($T_{60} = 0.25$ s, $\eta^2 = \infty$), ASE-TM achieves a PESQ of 2.96 and STOI of 0.99—closely matching the non-causal configuration.

6.5 DEREVERBERATION AND DECLIPPING ASE-TM MODEL ANALYSIS

Figure 4 presents the power spectra of the enhanced signals for the dereverberation and declipping tasks, over the entire test set. For both tasks, the spectrum of the enhanced signal $eh(n)$ exhibits significantly more power across a broad frequency range compared to the distorted input (reverberated or clipped after primary path), indicating successful signal restoration and enrichment. In the declipping task, it is particularly noteworthy that lower frequencies, crucial for speech intelligibility, show substantial power recovery in the enhanced signal’s spectrum. We further evaluated the declipping performance under a more aggressive clipping threshold of $\eta = 0.1$. The unprocessed clipped speech at this level yielded a PESQ score of 1.53 and an NMSE of -0.18 dB. ASE-TM restored these signals to a PESQ of 2.52 (CSIG 3.61, CBAK 2.76, COVL 3.08, STOI 0.91) and an NMSE of -1.22 dB. While these results are lower than for $\eta = 0.25$, they represent a substantial improvement over the severely clipped input, showing ASE-TM’s capability to handle extreme distortions.

7 CONCLUSIONS AND LIMITATIONS

In this paper, we introduced ASE, a novel paradigm that extends beyond traditional ANC by actively shaping the speech signal to enhance quality and intelligibility. Our ASE-TM model, which leverages a Transformer-Mamba architecture and a specialized loss function, demonstrated strong performance in denoising, dereverberation, and declipping, outperforming baseline methods. This study also reveals limitations that require further investigation. Baseline methods, designed for ANC, were adapted to ASE tasks, which may explain their reduced performance. Furthermore, future work should focus on developing a unified model that can handle multiple speech enhancement objectives, potentially leading to more versatile and efficient systems.

We utilized large language models (LLMs) to assist in refining the manuscript’s writing.

REFERENCES

- Jont B Allen and David A Berkley. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950, 1979.
- Christian Antoñanzas, Miguel Ferrer, Maria De Diego, and Alberto Gonzalez. Remote microphone technique for active noise control over distributed networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1522–1535, 2023.
- Jean-Eudes Ayilo, Mostafa Sadeghi, and Romain Serizel. Diffusion-based speech enhancement with a weighted generative-supervised learning loss. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12506–12510. IEEE, April 2024. doi: 10.1109/icassp48485.2024.10446805. URL <http://dx.doi.org/10.1109/ICASSP48485.2024.10446805>.
- Steven Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on acoustics, speech, and signal processing*, 27(2):113–120, 2003.
- Cassia Valentini Botinhao, Xin Wang, Shinji Takaki, and Junichi Yamagishi. Investigating rnn-based speech enhancement methods for noise-robust text-to-speech. In *9th ISCA speech synthesis workshop*, pp. 159–165, 2016.
- Ruizhe Cao, Sherif Abdulatif, and Bin Yang. Cmgan: Conformer-based metric gan for speech enhancement. *arXiv preprint arXiv:2203.15149*, 2022.
- Young-Jin Cha, Alireza Mostafavi, and Sukhpreet S Benipal. Dnoisenet: Deep learning-based feed-back active noise control in various noisy environments. *Engineering Applications of Artificial Intelligence*, 121:105971, 2023.
- Rong Chao, Wen-Huang Cheng, Moreno La Quatra, Sabato Marco Siniscalchi, Chao-Han Huck Yang, Szu-Wei Fu, and Yu Tsao. An investigation of incorporating mamba for speech enhancement. *arXiv preprint arXiv:2405.06573*, 2024.
- Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024.
- Danilo de Oliveira, Tal Peer, and Timo Gerkmann. Efficient transformer-based speech enhancement using long frames and stft magnitudes. In *Interspeech 2022*, pp. 2948–2952. ISCA, September 2022. doi: 10.21437/interspeech.2022-10781. URL <http://dx.doi.org/10.21437/Interspeech.2022-10781>.
- Alexandre Defossez, Gabriel Synnaeve, and Yossi Adi. Real time speech enhancement in the wave-form domain. *arXiv preprint arXiv:2006.12847*, 2020.
- Szu-Wei Fu, Chien-Feng Liao, Yu Tsao, and Shou-De Lin. Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement. In *International Conference on Machine Learning*, pp. 2031–2041. PmLR, 2019.
- Szu-Wei Fu, Cheng Yu, Tsun-An Hsieh, Peter Plantinga, Mirco Ravanelli, Xugang Lu, and Yu Tsao. Metricgan+: An improved version of metricgan for speech enhancement. In *Interspeech 2021*. ISCA, August 2021. doi: 10.21437/interspeech.2021-599. URL <http://dx.doi.org/10.21437/interspeech.2021-599>.
- Christopher C Fuller, Sharon Elliott, and Philip Arthur Nelson. *Active control of vibration*. Academic press, 1996.
- Sepehr Ghasemi, Raja Kamil, and Mohammad Hamiruce Marhaban. Nonlinear thf-fxlms algorithm for active noise control with loudspeaker nonlinearity. *Asian Journal of Control*, 18(2):502–513, 2016.
- Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

- Heitor R. Guimarães, Jiaqi Su, Rithesh Kumar, Tiago H. Falk, and Zeyu Jin. Ditse: High-fidelity generative speech enhancement via latent diffusion transformers, 2025.
- Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented transformer for speech recognition, 2020. URL <https://arxiv.org/abs/2005.08100>.
- Emanuel AP Habets. Room impulse response generator. *Technische Universiteit Eindhoven, Tech. Rep*, 2(2.4):1, 2006.
- Colin H Hansen, Scott D Snyder, Xiaojun Qiu, Laura A Brooks, and Danielle J Moreau. *Active control of noise and vibration*. E & Fn Spon London, 1997.
- Yanxin Hu, Yun Liu, Shubo Lv, Mengtao Xing, Shimin Zhang, Yihui Fu, Jian Wu, Bihong Zhang, and Lei Xie. Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement. *arXiv preprint arXiv:2008.00264*, 2020.
- Yi Hu and Philipos C Loizou. Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on audio, speech, and language processing*, 16(1):229–238, 2007.
- Hyung Yong Kim, Ji Won Yoon, Sung Jun Cheon, Woo Hyun Kang, and Nam Soo Kim. A multi-resolution approach to gan-based speech enhancement. *Applied Sciences*, 11(2):721, 2021.
- Sen M Kuo and Dennis R Morgan. Active noise control: a tutorial review. *Proceedings of the IEEE*, 87(6):943–973, 1999.
- Jean-Marie Lemerrier, Julius Richter, Simon Welker, and Timo Gerkmann. Storm: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2724–2737, 2023. ISSN 2329-9304. doi: 10.1109/taslp.2023.3294692. URL <http://dx.doi.org/10.1109/TASLP.2023.3294692>.
- Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, Itay Dalmedigos, Erez Safahi, Shaked Meirum, Yonatan Belinkov, Shai Shalev-Shwartz, et al. Jamba: A hybrid transformer-mamba language model. *arXiv preprint arXiv:2403.19887*, 2024.
- Jae Lim and Alan Oppenheim. All-pole modeling of degraded speech. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(3):197–210, 1978.
- Haohe Liu, Xubo Liu, Qiuqiang Kong, Qiao Tian, Yan Zhao, DeLiang Wang, Chuanzeng Huang, and Yuxuan Wang. Voicefixer: A unified framework for high-fidelity speech restoration. *arXiv preprint arXiv:2204.05841*, 2022.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- Ye-Xin Lu, Yang Ai, and Zhen-Hua Ling. Mp-senet: A speech enhancement model with parallel denoising of magnitude and phase spectra. *arXiv preprint arXiv:2305.13686*, 2023.
- Yen-Ju Lu, Zhong-Qiu Wang, Shinji Watanabe, Alexander Richard, Cheng Yu, and Yu Tsao. Conditional diffusion probabilistic model for speech enhancement. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7402–7406. Ieee, 2022.
- Paul Lueg. Process of silencing sound oscillations. *US patent 2043416*, 1936.
- Zhengding Luo, Dongyuan Shi, and Woon-Seng Gan. A hybrid sfanc-fxnlms algorithm for active noise control based on deep learning. *IEEE Signal Processing Letters*, 29:1102–1106, 2022.
- Zhengding Luo, Dongyuan Shi, Xiaoyi Shen, and Woon-Seng Gan. Unsupervised learning based end-to-end delayless generative fixed-filter active noise control. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 441–445. IEEE, 2024.

- Yehuda Mishaly, Lior Wolf, and Eliya Nachmani. Deep active speech cancellation with multi-band mamba network, 2025. URL <https://arxiv.org/abs/2502.01185>.
- Alireza Mostafavi and Young-Jin Cha. Deep learning-based active noise control on construction sites. *Automation in Construction*, 151:104885, 2023.
- Philip Arthur Nelson and Stephen J Elliott. *Active control of sound*. Academic press, 1991.
- Kuldip Paliwal, Belinda Schwerin, and Kamil Wójcicki. Speech enhancement using a minimum mean-square error short-time spectral modulation magnitude estimator. *Speech Communication*, 54(2):282–305, 2012.
- Ashutosh Pandey and DeLiang Wang. A new framework for supervised speech enhancement in the time domain. In *Interspeech*, pp. 1136–1140, 2018.
- JungPhil Park, Jeong-Hwan Choi, Yungyeo Kim, and Joon-Hyuk Chang. Had-anc: A hybrid system comprising an adaptive filter and deep neural networks for active noise control. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2023, pp. 2513–2517. International Speech Communication Association, 2023.
- Santiago Pascual, Antonio Bonafonte, and Joan Serra. Segan: Speech enhancement generative adversarial network. *arXiv preprint arXiv:1703.09452*, 2017.
- Alexander Pike and Jordan Cheer. Generalized performance of neural network controllers for feed-forward active control of nonlinear systems. 2023.
- Mirco Ravanelli, Titouan Parcollet, Adel Moumen, Sylvain de Langen, Cem Subakan, Peter Plantinga, Yingzhi Wang, Pooneh Mousavi, Luca Della Libera, Artem Ploujnikov, et al. Open-source conversational ai with speechbrain 1.0. *Journal of Machine Learning Research*, 25(333): 1–11, 2024.
- Xing Ren and Hongwei Zhang. An improved artificial bee colony algorithm for model-free active noise control: algorithm and implementation. *IEEE Transactions on Instrumentation and Measurement*, 71:1–11, 2022.
- Dario Rethage, Jordi Pons, and Xavier Serra. A wavenet for speech denoising. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5069–5073. IEEE, 2018.
- 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*, 31:2351–2364, 2023. ISSN 2329-9304. doi: 10.1109/taslp.2023.3285241. URL <http://dx.doi.org/10.1109/TASLP.2023.3285241>.
- Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. 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)*, volume 2, pp. 749–752. IEEE, 2001.
- Shrishti Saha Shetu, Emanuël AP Habets, and Andreas Brendel. Gan-based speech enhancement for low snr using latent feature conditioning. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.
- Dongyuan Shi, Bhan Lam, Kenneth Ooi, Xiaoyi Shen, and Woon-Seng Gan. Selective fixed-filter active noise control based on convolutional neural network. *Signal Processing*, 190:108317, 2022.
- Dongyuan Shi, Bhan Lam, Xiaoyi Shen, and Woon-Seng Gan. Multichannel two-gradient direction filtered reference least mean square algorithm for output-constrained multichannel active noise control. *Signal Processing*, 207:108938, 2023.
- Dongyuan Shi, Woon-seng Gan, Xiaoyi Shen, Zhengding Luo, and Junwei Ji. What is behind the meta-learning initialization of adaptive filter?—a naive method for accelerating convergence of adaptive multichannel active noise control. *Neural Networks*, 172:106145, 2024.

- Wooseok Shin, Byung Hoon Lee, Jin Sob Kim, Hyun Joon Park, and Sung Won Han. Metricgan-okd: multi-metric optimization of metricgan via online knowledge distillation for speech enhancement. In *International Conference on Machine Learning*, pp. 31521–31538. PMLR, 2023.
- Deepali Singh, Rinki Gupta, Arun Kumar, and Rajendar Bahl. Enhancing active noise control through stacked autoencoders: Training strategies, comparative analysis, and evaluation with practical setup. *Engineering Applications of Artificial Intelligence*, 135:108811, 2024.
- Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong. Attention is all you need in speech separation, 2021. URL <https://arxiv.org/abs/2010.13154>.
- Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE international conference on acoustics, speech and signal processing*, pp. 4214–4217. IEEE, 2010.
- Wenxin Tai, Fan Zhou, Goce Trajcevski, and Ting Zhong. Revisiting denoising diffusion probabilistic models for speech enhancement: Condition collapse, efficiency and refinement. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 13627–13635, 2023.
- Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent. The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings. In *Proceedings of Meetings on Acoustics*, volume 19. AIP Publishing, 2013.
- Christophe Veaux, Junichi Yamagishi, and Simon King. The voice bank corpus: Design, collection and data analysis of a large regional accent speech database. In *2013 international conference oriental COCOSDA held jointly with 2013 conference on Asian spoken language research and evaluation (O-COCOSDA/CASLRE)*, pp. 1–4. IEEE, 2013.
- Kai Wang, Bengbeng He, and Wei-Ping Zhu. Tstnn: Two-stage transformer based neural network for speech enhancement in the time domain. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7098–7102. IEEE, June 2021. doi: 10.1109/icassp39728.2021.9413740. URL <http://dx.doi.org/10.1109/ICASSP39728.2021.9413740>.
- Simon Welker, Julius Richter, and Timo Gerkmann. Speech enhancement with score-based generative models in the complex stft domain. *arXiv preprint arXiv:2203.17004*, 2022.
- Tong Xiao, Buye Xu, and Chuming Zhao. Spatially selective active noise control systems. *The Journal of the Acoustical Society of America*, 153(5):2733–2733, 2023.
- Yang Xiao and Rohan Kumar Das. Tf-mamba: A time-frequency network for sound source localization. *arXiv preprint arXiv:2409.05034*, 2024.
- Moujia Ye and Hongjie Wan. Improved transformer-based dual-path network with amplitude and complex domain feature fusion for speech enhancement. *Entropy*, 25(2):228, 2023.
- Hao Zhang and DeLiang Wang. Deep anc: A deep learning approach to active noise control. *Neural Networks*, 141:1–10, 2021.
- Hao Zhang and DeLiang Wang. Deep mcanc: A deep learning approach to multi-channel active noise control. *Neural Networks*, 158:318–327, 2023.
- Hao Zhang, Ashutosh Pandey, and DeLiang Wang. Attentive recurrent network for low-latency active noise control. In *INTERSPEECH*, pp. 956–960, 2022a.
- Hao Zhang, Ashutosh Pandey, et al. Low-latency active noise control using attentive recurrent network. *IEEE/ACM transactions on audio, speech, and language processing*, 31:1114–1123, 2023a.
- Huawei Zhang, Jihui Zhang, Fei Ma, Prasanga N Samarasinghe, and Huiyuan Sun. A time-domain multi-channel directional active noise control system. In *2023 31st European Signal Processing Conference (EUSIPCO)*, pp. 376–380. IEEE, 2023b.

- Qiquan Zhang, Hongxu Zhu, Xinyuan Qian, Eliathamby Ambikairajah, and Haizhou Li. An exploration of length generalization in transformer-based speech enhancement. In *Interspeech 2024*, pp. 1725–1729. ISCA, September 2024. doi: 10.21437/interspeech.2024-1831. URL <http://dx.doi.org/10.21437/interspeech.2024-1831>.
- Shucong Zhang, Malcolm Chadwick, Alberto Gil CP Ramos, and Sourav Bhattacharya. Cross-attention is all you need: Real-time streaming transformers for personalised speech enhancement. *arXiv preprint arXiv:2211.04346*, 2022b.
- Yang Zhou, Haiquan Zhao, and Dongxu Liu. Genetic algorithm-based adaptive active noise control without secondary path identification. *IEEE Transactions on Instrumentation and Measurement*, 2023.