FSD: Acoustic Echo Cancellation with Fewer Step Diffusion

Yang Liu
Meta, US
yangliuai@meta.com

Li Wan
Meta, US
wwanli@meta.com

Yiteng Huang
Meta, US
yah@meta.com

Ming Sun
Meta, US
sunming425@meta.com

Changsheng Zhao
Meta, US
cszhao@meta.com

Zhaoheng Ni
Meta, US
zni@meta.com

Xinhao Mei
Meta, US
xinhaomei@meta.com

Yangyang Shi
Meta, US
yyshi@meta.com

Florian Metze
Meta, US
fmetze@meta.com

Abstract

Despite the promising capabilities of diffusion models in speech enhancement, their application in Acoustic Echo Cancellation (AEC) has been limited. In this paper, we introduce Fewer Step Diffusion, a framework specifically designed for AEC, which addresses computational efficiency concerns, making it particularly suitable for deployment on edge devices. Unlike traditional approaches, FSD uses a novel score model, which substantially boosts processing efficiency. Additionally, we present a unique noise generation technique that leverages far-end signals, utilizing both far-end and near-end signals to enhance the accuracy of the score model. We evaluate our proposed method using the ICASSP2023 Microsoft Deep Echo Cancellation Challenge dataset, where FSD demonstrates superior performance compared to several end-to-end methods and other diffusion-based echo cancellation techniques.

1 Introduction

The importance of acoustic echo cancellation (AEC) in achieving high-quality speech during voice communication has led to the development of various methods, including adaptive filtering techniques such as Least Mean Squares (LMS) (1) and Recursive Least Squares (RLS) (2), and deep neural network (DNN) based approaches like the Deep Complex Convolution Recurrent Network (DC-CRN) (3). The main challenges in AEC are artifacts, target speech distortion, and echo leakage, especially during double-talk scenarios. To address these issues, researchers have explored solutions such as alignment modules (4), novel architectures (5; 6), and modified loss functions (7).

Among the recent advancements, diffusion-based generative models have shown promising results in tasks like noise suppression. These models generate data by reversing a gradual noise process and consist of two main phases: the forward diffusion process and the reverse generative process. For example, Lu et al. (8) proposed an approach that integrates the characteristics of noisy speech into both diffusion and reverse phases, resulting in a more refined enhancement. Similarly, Joan Serrà et al. (9) introduced a multi-resolution conditioning network employing score-based diffusion, which generates clean speech by progressively reducing noise in a series of steps. Lemercier et al. (10) applied a stochastic regeneration strategy, using estimates from a predictive model as guides for further diffusion, thereby refining the enhancement process.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

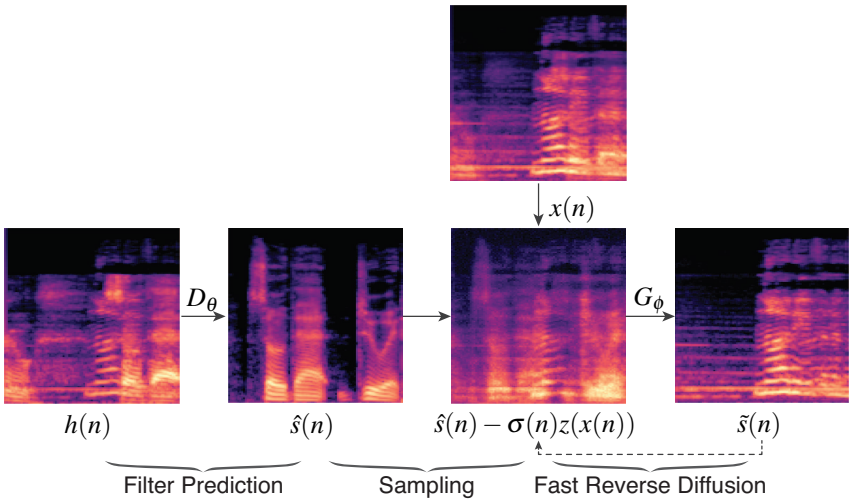

Figure 1: FSD pipeline. The predictive filter first generates an estimate $\hat{\mathbf{s}}(n)$ from the microphone signal $\mathbf{h}(n)$. The diffusion-based generation module G_ϕ then adds Gaussian noise, guided by the far-end signal $\mathbf{x}(n)$, and solves the reverse diffusion stochastic differential equation (SDE). The resulting estimated near-end speech $\tilde{\mathbf{s}}(n)$ is used in the score function for the next frame.

While there have been notable advances in speech enhancement, the use of diffusion models specifically for echo cancellation remains underexplored. The primary challenge is the computational intensity associated with traditional diffusion models, which poses difficulties for real-time deployment. In this paper, we propose a novel model, FSD (Fewer Step Diffusion), which utilizes a diffusion-based framework tailored for AEC. FSD is designed to overcome the computational barriers by executing the diffusion process with fewer steps, running the score model only once per frame and using prior states to reduce processing time significantly. Furthermore, FSD incorporates a novel noise generation technique that utilizes both far-end and near-end signals to enhance accuracy, leading to improved echo cancellation performance. Our experiments demonstrate that FSD achieves superior results on the ICASSP2023 Microsoft Deep Echo Cancellation Challenge dataset compared to existing end-to-end and diffusion-based methods.

2 Proposed Method

2.1 Problem Formulation

In a typical AEC system, the microphone signal is denoted as $\mathbf{h}(n)$. This signal comprises two components: the near-end speech $\mathbf{s}(n)$ and the acoustic echo $\mathbf{z}(n)$. Mathematically, this relationship is expressed as:

$$\mathbf{h}(n) = \mathbf{s}(n) + \mathbf{z}(n), \tag{1}$$

where n is the time sample index. The acoustic echo $\mathbf{z}(n)$ can be understood as a time-delayed version of the far-end reference signal $\mathbf{x}(n)$. This signal has traversed the echo path and might have undergone nonlinear distortions due to the loudspeakers. The primary objective of the AEC system is to separate the near-end speech $\mathbf{s}(n)$ from the microphone signal $\mathbf{h}(n)$.

2.2 Forward Process and Inference through Reverse Sampling

The perturbation kernel plays a crucial role in the diffusion process by introducing Gaussian noise into the data, which helps regularize the model, preventing overfitting and ensuring robust performance. The forward process is defined by the Itô Stochastic Differential Equation (SDE), describing how the data evolves over time with added noise. The score matching objective is essential for training the score model. It ensures that the model's predictions match the true data gradients, improving

the accuracy of the denoised outputs. By minimizing the score matching loss, we train the score model to accurately capture the underlying data distribution.

The stochastic forward process utilized in score-based diffusion models is defined by the Itô Stochastic Differential Equation (SDE) (11):

$$\mathrm{d}\mathbf{s}(n)_t = \mathrm{f}(\mathbf{s}(n)_t, t)\mathrm{d}t + \mathrm{g}(t)\mathrm{d}\mathbf{w} \tag{2}$$

where $\mathbf{w}$ denotes a standard-dimensional Brownian motion, making $\mathrm{d}\mathbf{w}$ a zero-mean Gaussian random variable with variance proportional to $\mathrm{d}t$, pertinent for each Time-Frequency (T-F) bin. The functions f and g represent the drift and diffusion coefficients, respectively. The state of the process at discrete index n and continuous time t, where $t \in [0, T]$, is given by $\mathbf{s}(n)_t$, and for clean speech, the initial condition is $\mathbf{s}(n)_0 = \mathbf{s}(n)$.

In the reverse process of the score-based diffusion model, the score model is substituted into the reverse SDE as a plug-in reverse SDE (12):

$$\mathrm{d}\mathbf{s}(n)_t = [-\mathrm{f}(\mathbf{s}(n)t, t) + \mathrm{g}(t)^2 \nabla \mathbf{s}(n)_t \log p_t(\mathbf{s}(n)_t)]\mathrm{d}t + \mathrm{g}(t)\mathrm{d}\overline{\mathbf{w}}, \tag{3}$$

where $\mathrm{d}\overline{\mathbf{w}}$ is a d-dimensional Brownian motion for the time flowing in reverse, and $\nabla_{\mathbf{s}(n)_t} \log p_t(\mathbf{s}(n)_t)$ is the score function. This equation is classified under the Ornstein-Uhlenbeck SDEs (13).

During inference, Eq. (3) is evaluated using the predictor-corrector approach informed by the score-matching network described in (11). The initial state of the process is drawn from the distribution:

$$\mathbf{s}(n)\tau \sim \mathscr{N}\mathbb{C}(\mathbf{s}(n)\tau; \mathbf{h}(n), \mathbf{x}(n), \sigma^2(\tau)\mathbf{I}), \tag{4}$$

which represents a near-end signal $\mathbf{h}(n)$ and far-end signal $\mathbf{x}(n)$, with Gaussian noise of variance $\sigma^2(\tau)$ added.

2.3 Fewer Step Score Model with Far-End Guided Noise

Since the speech enhancement task, including AEC, can be considered a conditional generation task, the conditioning is integrated into the diffusion process. The forward process yields a complex Gaussian distribution for the process state $\mathbf{s}(n)_t$, known as the perturbation kernel (14):

$$\mathscr{N}_{\mathbb{C}}(\mathbf{s}(n)_t; \boldsymbol{\mu}(\mathbf{s}(n)_0, \mathbf{h}(n), t), \sigma(t)^2\mathbf{I}), \tag{5}$$

where the mean is $\boldsymbol{\mu}$ and the variance is $\sigma(t)^2$. During inference, one attempts to solve the reverse SDE in Eq. (3). For the Gaussian form of the perturbation kernel $p_{0,t}(\mathbf{s}(n)_t|\mathbf{s}(n)0, \mathbf{h}(n), \mathbf{x}(n))$ for the AEC task, and the regularity conditions exhibited by the mean and variance, a score matching objective is used to train the score model $\mathbf{s}\phi$. This score model, adapted from the StoRM architecture (10), is a neural network designed to estimate the gradient of the data distribution.

2.4 Fewer Step Diffusion (FSD) Model

To address the computational challenges of deploying full diffusion-based models in real-time, we propose the **FSD (Fewer Step Diffusion)** model, which significantly reduces the number of diffusion steps required for echo cancellation. The FSD model is designed to execute the score calculation only once per frame, leveraging the enhanced signal from the previous frame to reduce computational overhead. This approach balances the trade-off between accuracy and processing time, making it suitable for deployment on edge devices.

The score function of the perturbation kernel is:

$$\mathscr{J}^{(\mathrm{DSM})}(\boldsymbol{\phi}) = \mathbb{E}[|\hat{\mathbf{s}}_{\boldsymbol{\phi}}(\mathbf{s}(n), \mathbf{h}(n), s(n-1)) + \frac{\mathbf{z}}{\sigma(n)}|_2^2], \tag{6}$$

where $\mathscr{J}^{(\mathrm{DSM})}$ is related to the enhanced signal from the previous frame rather than the time index t. This simplification reduces computational demands for each frame by optimizing over time with fewer score calculations.

2.5 Loss Function

For training, we define the loss $\mathscr{L}$ as a combination of score matching and a supervised regularization term — e.g., mean square error — matching the output of the initial predictor to the target speech:

$$\begin{aligned}\mathscr{L}^{(\mathrm{StoRM})}(\theta,\phi) &= \mathscr{L}^{(\mathrm{DSM})}(\theta) + \alpha\mathscr{L}^{(\mathrm{Sup})}(\phi) \\ &= \mathbb{E}\left[\left|\mathbf{s}_\phi(\hat{\mathbf{s}}(n),\mathbf{h}(n),s(n-1)) + \frac{\mathbf{z}}{\sigma(n)}\right|_2^2\right] + \alpha\mathbb{E}\left[|\mathbf{s}(n) - D_\theta(\mathbf{h}(n))|_2^2\right],\end{aligned} \tag{7}$$

where α is a balancing term empirically set to 1.

3 Experimentation Results

3.1 Data Selection and Training

For model training, we utilize a combination of synthetic data from the AEC-challenge (15) and our privately enhanced dataset. We ensure gender balance among speakers on both the far-end and near-end sides, resulting in 720 original conversations, each lasting 10 seconds.

The training configuration for the model involves using the Adam optimizer with an initial learning rate of 10^{-4}. The model is trained for 160 epochs with an effective batch size of 32, distributed across multiple GPUs using the distributed data-parallel approach in PyTorch Lightning. During training, an exponential moving average of the model weights is tracked with a decay rate of 0.999, which is used for sampling. The training steps include sampling a random time t, sampling clean and noisy speech pairs from the dataset, adding noise to the clean speech according to the forward diffusion process, and computing the loss as an L2 loss between the model output and the score of the perturbation kernel. Hyperparameter choices for the diffusion process include $\sigma_{\min} = 0.05$, $\sigma_{\max} = 0.5$, and $\gamma = 1.5$, selected based on empirical hyperparameter optimization. For the spectrogram transformation, $\alpha = 0.5$ and $\beta = 0.15$ were chosen. Hyperparameter tuning was conducted through grid search, adjusting the number of reverse steps N and the step size parameter r for the annealed Langevin corrector to balance performance and computational efficiency.

3.2 Ablation Study

Table 1 presents the performance of various acoustic echo cancellation (AEC) models based on different sampling methods and the presence of reverse diffusion. The models are evaluated on metrics such as ERLE of FEST, PESQ of NEST, and PESQ of DT.

Initially, we observe the performance of the CRN model (3) without any sampling or reverse diffusion. Specifically, the larger model with higher latency shows superior performance in both ERLE of FEST and PESQ of NEST, indicating that increasing the model size can enhance AEC performance, given similar architectural characteristics. The diffusion-based models (IDs 3 and 4) introduce reverse diffusion with distinct sampling methods — Random Sampling and Far-End Guided Sampling, respectively. These models exhibit a high latency of 325.00 ms but surpass the CRN models in all performance metrics due to the added diffusion steps.

Our proposed FSD (Fewer Step Diffusion) model combines a modified reverse approach, termed "Fast Reversion," with sampling techniques. The FSD model achieves a latency of 9.14 ms, which is well within the real-time processing requirement (typically 10 ms). This low latency is achieved by running the score model only once per frame, significantly reducing the computational load compared to traditional diffusion-based methods, thereby avoiding perceptible delays and maintaining the quality and naturalness of the conversation. Performance-wise, the FSD model with Far-End Guided Sampling marginally outperforms its Random Sampling counterpart, particularly in ERLE of FEST and PESQ of DT.

Model	CRN	CRN	Diffusion-Based Models	
Sampling	No	No	Random / Far-End Guided	Random / Far-End Guided
Reversion Diffusion	No	No	Yes	Fewer Step Score
Parameters (M)	3.6	7.8	6.9	6.9
Latency (ms)	4.04	8.93	325.00	9.14
ERLE of FEST (dB)	67.67	82.45	92.51 / 92.83	85.80 / 89.75
PESQ of NEST	4.41	4.50	4.87 / 4.91	4.79 / 4.85
PESQ of DT	2.34	2.60	3.30 / 3.32	3.05 / 3.24

Table 1: Performance comparison over candidate models. We measure PESQ for both DT and NEST scenarios and ERLE for the FEST scenario in the augmented evaluation dataset. Slight improvements or declines in the results for different sampling methods are noted in the diffusion-based models.

Model	FEST	DT	DT other
RLS	2.64	2.47	3.78
ByteAudio	4.709	4.770	4.312
FSD	4.830	4.820	4.470

Table 2: AECMOS comparison on ICASSP 2023 AEC Challenge blind test.

To summarize, using the diffusion-based model, the DT PESQ improved by 27.7%, rising from 2.6 to 3.32. When implementing the FSD model, the DT PESQ slightly decreased to 3.08, representing a 7.2% reduction. However, a significant advantage of the FSD model is its reduced latency, which is only 2.8% of the diffusion-based model's latency (9.14 ms vs. 325 ms).

3.3 Comparison with State-of-the-Art Methods

We use AECMOS, a non-intrusive model-based metric from the AEC challenge, to compare our proposed method against established baselines such as Recursive Least Squares (RLS) (2), ByteAudio (16), and diffusion-based models. ByteAudio employs a Two-step Band-split Neural Network (TBNN) methodology for full-band acoustic echo cancellation, achieving the highest performance in the AEC 2023 challenges, second only to the host.

The results show that our proposed models outperform RLS in both single talk and double talk scenarios. As shown in Table 2, the diffusion-based model slightly outperforms ByteAudio across all three scenarios. All four methods demonstrate notable performance on the AEC task, but the FSD model balances acceptable performance with low latency, reducing computational load while potentially missing some nuances. These results underscore the importance of selecting the appropriate AEC method for specific tasks and scenarios and highlight the progress and potential of contemporary AEC technologies.

4 Conclusion

We proposed fewer step diffusion, a novel score-based diffusion model specifically designed for AEC. This research demonstrates that diffusion-based stochastic regeneration models can significantly enhance AEC performance. To address the computational cost challenges associated with traditional diffusion models, particularly for edge devices, FSD improves processing efficiency and reduces latency by running the score model only fewer times per frame. Additionally, the FSD model uses noise generation guided by far-end signals, incorporating both far-end and near-end signals to enhance the precision of the score model. This unique application of diffusion models offers a powerful and efficient approach to echo cancellation.

In future work, we will focus on deploying the FSD model in real-time applications, which involves potential challenges such as ensuring adequate computational resources, managing scalability for multiple streams, and exploring the integration of our methods with other advanced audio processing techniques, such as noise suppression and dereverberation.

References

[1] Bernard Widrow and ME Hoff, "Ire wescon convention record," *IRE, New York*, pp. 96–104, 1960.

[2] Simon S Haykin, *Adaptive filter theory*, Pearson Education India, 2002.

[3] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Proc. InterSpeech*, 2021, p. 2472–2476.

[4] Y. Liu, Y. Shi, Y. Li, K. Kalgaonkar, S. Srinivasan, and X. Lei, "SCA: Streaming cross-attention alignment for echo cancellation," in *Proc. IEEE ICASSP*. IEEE, 2023, pp. 1–5.

[5] S. Zhang, Y. Kong, S. Lv, Y. Hu, and L. Xie, "FT-LSTM based complex network for joint acoustic echo cancellation and speech enhancement," *arXiv preprint arXiv:2106.07577*, 2021.

[6] H. Zhao, N. Li, R. Han, L. Chen, X. Zheng, C. Zhang, L. Guo, and B. Yu, "A deep hierarchical fusion network for fullband acoustic echo cancellation," in *Proc. IEEE ICASSP*. IEEE, 2022, pp. 9112–9116.

[7] F. Xiong, M. Dong, K. Zhou, H. Zhu, and J. Feng, "Deep subband network for joint suppression of echo, noise and reverberation in real-time fullband speech communication," in *Proc. IEEE ICASSP*. IEEE, 2023, pp. 1–5.

[8] Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao, "Conditional diffusion probabilistic model for speech enhancement," in *Proc. IEEE ICASSP*. IEEE, 2022, pp. 7402–7406.

[9] J. Serrà, S. Pascual, J. Pons, R.O. Araz, and D. Scaini, "Universal speech enhancement with score-based diffusion," *arXiv preprint arXiv:2206.03065*, 2022.

[10] J.-M. Lemercier, J. Richter, S. Welker, and T. Gerkmann, "StoRM: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation," *arXiv preprint arXiv:2212.11851*, 2022.

[11] Y. Song, J. Sohl-Dickstein, D.P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2020.

[12] C.-W. Huang, J.H. Lim, and A.C. Courville, "A variational perspective on diffusion-based generative models and score matching," *Advances in Neural Information Processing Systems*, vol. 34, pp. 22863–22876, 2021.

[13] B. Oksendal, *Stochastic differential equations: an introduction with applications*, Springer Science and Business Media, 2013.

[14] S. Särkkä and A. Solin, *Applied stochastic differential equations*, vol. 10, Cambridge University Press, 2019.

[15] R. Cutler, A. Saabas, T. Parnamaa, M. Purin, H. Gamper, S. Braun, K. Sorensen, and R. Aichner, "ICASSP 2022 acoustic echo cancellation challenge," in *Proc. IEEE ICASSP*, 2022.

[16] Z. Zhang, S. Zhang, M. Liu, Y. Leng, Z. Han, L. Chen, and L. Xie, "Two-step band-split neural network approach for full-band residual echo suppression," in *Proc. IEEE ICASSP*. IEEE, 2023, pp. 1–2.