2025 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) | 979-8-3315-3745-6/25/$31.00 ©2025 IEEE | DOI: 10.1109/WASPAA66052.2025.11230936

A Unified Framework for Evaluating DNN-Based Feedforward, Feedback, and Hybrid Active Noise Cancellation

Huajian Fang[1*], *Buye Xu*[2], *Jacob Donley*[2], *Ashutosh Pandey*[2], *Deliang Wang*[3], *Daniel Wong*[2]

[1]University of Hamburg, Germany [2]Meta Reality Labs, USA [3] Ohio State University, USA

***Abstract*—Deep neural network (DNN)-based acoustic noise cancellation excels at modeling complex non-linear relationships in signal patterns that are difficult for linear filters to handle. Recent work has studied DNN-based feedforward (FF) and feedback (FB) control structures. However, their inconsistent experimental settings prevent fair comparison, and limited evaluation in a narrow range of acoustic conditions hinders a comprehensive understanding of their effectiveness. In this work, we present a unified framework that realizes FF and FB control structures to better understand their cancellation mechanisms under various acoustic conditions. In addition, it enables a hybrid (HB) control structure that combines the FF and FB approaches, which has not been evaluated against standalone DNN-based FF and FB configurations. We perform systematic evaluations in various settings, including different disturbing signals, reverberation conditions, source positions, room sizes, and mismatched secondary paths. The evaluation results show that the effectiveness of FF depends strongly on the modeling of the primary path and the acoustic environment, the performance of FB varies less under different acoustic conditions, and HB integrates the advantages of FF and FB.**

1. INTRODUCTION

Acoustic noise cancellation (ANC) aims to attenuate unwanted ambient noise at a cancellation point by playing an anti-noise with the same amplitude but the opposite phase [1]–[4]. ANC is widely used in close-ear headphones, where an ANC algorithm addresses narrowband low-frequency attenuation, while ear cushions passively attenuate high-frequency components [5], [6]. However, emerging devices such as virtual reality (VR) headsets may necessitate open-ear ANC, which requires effective broadband frequency attenuation. This work aims to provide insights into the challenges involved in open-ear ANC.

Depending on how the disturbance is acquired, the active cancellation of unwanted sounds can be achieved with different control structures [2], [4], [7]: feedforward (FF), feedback (FB), and hybrid (HB). The FF control system generates an anti-noise based on primary noise captured by a reference microphone. The FB control system relies solely on residue error signals recorded by an error microphone. Additionally, the HB control structure makes use of both the reference and error microphone recordings.

A widely used adaptive algorithm for implementing these control structures is the filtered-x least mean square (FxLMS) algorithm and its variants [2], [4], [8], [9]. In contrast to adaptive filtering, which adjusts in real-time but has complexity and convergence issues, the fixed-filter method pre-learns a set of control filters during training and then selects the most appropriate filter to attenuate sounds during operation [10]. Recent work has further improved the FF fixed-filter method by considering the filter selection as a classification task and utilizing a deep neural network (DNN) as a classifier [11]. Luo et al. [12] proposed to further fine-tune the selected filters during operation. While these DNN-assisted approaches combining conventional signal processing algorithms with deep learning techniques demonstrate improved performance [11]–[13], they still inherit the limitations of traditional signal processing algorithms, such as their linearity, stability issues, and difficulty in handling non-stationary sounds [14], [15]. Thus, performance degradation may be observed when attenuating real-world ambient noise under open-ear conditions, as such noise typically features non-stationary and broadband characteristics.

In addition to DNN-assisted methods, fully DNN-based methods have been proposed to learn a control module in an end-to-end manner. Zhang et al. formulated the FF control system as an end-to-end supervised learning problem in the short-time Fourier transform (STFT) domain [16]. Further extensions include the use of a more advanced architecture [17] and multichannel reference microphones [18]. DNN-based FB algorithms that perform in the waveform domain have also been proposed in [19], where the recursive feedback loop is modeled with a recurrent neural network. A similar principle of using a DNN model to replace the FB filtering module has been studied in [20], [21]. Despite the high performance reported in the literature, little is known about their robustness under various acoustic conditions. Existing analyses are often limited to a narrow range of acoustic environments (e.g., fixed noise positions, room sizes, and acoustic settings [16], [19]) and do not provide a full picture of their effectiveness. Yet, gaining such understanding is essential to fully leverage the modeling potential of DNNs. Moreover, independent development of FF and FB control structures has led to variations in DNN architectures, training strategies, and experimental setups. For instance, existing work on DNN-based FF systems [16]–[18] is often implemented as frame-based algorithms, whereas DNN-based FB methods [19]–[21] often operate on a sample-by-sample basis. As it is not fair to compare these approaches directly, their relative performance remains unclear.

To address these limitations, we present a unified framework that facilitates comprehensive evaluation and fair comparison. The unified framework can achieve FB and FF topologies and operates in a frame-based causal manner. In addition, the framework implements the HB control structure, which, to the authors' knowledge, has not been explored in DNN-based studies. Rather than restricting the evaluation to a limited number of configurations [16], [17], [19], our analysis extensively examines the impact of acoustic simulation techniques (such as the image source (IS) and ray tracing (RT) methods [22], [23]) and environmental variations (such as disturbing source types, reverberation conditions, source positions, and room sizes) on their performance. Further studies examine their stability to acoustic changes in the secondary path between training and testing. The experimental results provide insights into the effectiveness and robustness of ANC control structures across diverse acoustic conditions.

2. PROBLEM STATEMENT

In this work, we consider an ANC system that attenuates an unwanted signal $d(t)$ at the cancellation point by emitting a compensation signal $y(t)$ through a loudspeaker. t indexes samples in the time domain. The control signal $y(t)$ propagates through an acoustic path $S(z)$, also called a secondary path, reaches the cancellation point as an anti-sound $y'(t)$, and destructively superimposes with $d(t)$: $e(t) = d(t) - y'(t)$. $e(t)$ is the residual error signal and is recorded by the error microphone placed at the cancellation point. $S(z)$ denotes the z-transform of $s(t)$.

* Work done during the internship at Meta.

In a FF control system, the control signal $y(t)$ can be generated by an ANC module filtering the signals $x(t)$ recorded by a reference microphone. This ANC module, denoted as $f_{\text{FF}}(\cdot)$, can usually be realized by a FxLMS-based adaptive filtering method. The effectiveness of the FF system is often related to the coherence of the disturbance at the reference and error microphones. $f_{\text{FF}}(\cdot)$ is optimized to model the primary path (an acoustic path between the reference and error microphones) while also considering the secondary path. Additionally, it is possible to leverage spatial information when multiple reference microphones are available. In this setup, we consider multichannel reference signals, as will be described later in Section 3.1. In contrast, the FB system relies only on the error microphone and recreates the disturbance at the cancellation point with an estimated secondary path $S'(z)$:

$$d'(t) = e(t) + y(t) * s'(t), \tag{1}$$

where $*$ denotes the linear convolution. In traditional FxLMS-based FB algorithms, the regenerated signal $d'(t)$ can be used by a FB ANC control module, $f_{\text{FB}}(\cdot)$, in the same way as the reference signal in the FF system. Finally, the HB system combines the FF and FB principles and generates control signals using both the disturbance at the reference microphone and the residue signal at the error microphone.

3. A UNIFIED FRAMEWORK FOR DNN-BASED ANC

We present a unified DNN-based framework that can realize FF and FB control structures, as well as their combination in a HB system.

3.1. DNN-based FF Control Structure

The DNN-based FF system can be implemented by replacing the traditional filtering module with a DNN, denoted as $f_{\text{FF}}(\cdot)$. By collecting or simulating paired data from the reference and error microphones, we can formulate the FF system as a supervised learning problem, following [16]. By incorporating the secondary path into the forward process, the DNN can be optimized using a loss function like mean squared error (MSE). We model the signal in frames rather than processing it sample by sample to avoid the computational complexity caused by running a DNN frequently, which is specifically expensive at high sampling rates. However, one disadvantage of block-based ANC is the delay caused by the frame length [24]. This can be alleviated to some extent by incorporating a prediction module with parametric signal processing techniques [25], [26]. Given the high complexity of real-world signals such as speech and environmental noise, a data-driven DNN offers a more powerful alternative for the prediction of future content. Thus, we use a DNN to achieve not only control signal generation, but also future prediction, i.e., predicting a control signal for the n-th frame by processing only the reference signal up to the $n-1$-th frame.

Following the success of STFT representations in speech and audio signal processing [27], [28], the DNN can take as input the STFT representations of reference microphone signals and output time-frequency representations of control signals [16]. This is followed by the standard inverse Fourier transform and overlap-add procedures. For STFT representations, the window length is typically set to balance the time and frequency resolution, e.g., 20 ms. However, this window length poses challenges for DNNs attempting to make frame-based predictions. To remedy this limitation, we use a shorter window for synthesis while maintaining a longer analysis window. This enables DNNs to predict short frames without compromising frequency resolution.

Modern devices like smart glasses or headsets are often equipped with multiple microphones. We assume a reference microphone array surrounding the cancellation point, where the cancellation point can be analogized to the inside of the ear canal. The microphone array provides a multichannel reference signal $\mathbf{X}_n = \{\mathbf{x}_n^c\}_{c=1}^{c=C}$, adding spatial information compared to the single-channel case. C denotes the number of reference microphones and $\mathbf{x}_n = \{x_n(m)\}_{m=1}^{m=M}$ denotes the n-th frame signal in the time domain, where m indexes the samples within a frame and M can be set according to the window length of the STFT. Finally, the FF system can output the n-th frame control signal $\mathbf{y}_n$ by processing the reference signal of the preceding frames:

$$\mathbf{y}_n = f_{\text{FF}}(\mathbf{X}_{n-1}, \cdots, \mathbf{X}_1). \tag{2}$$

Given aligned STFT representations of the reference and error microphones, future prediction can be achieved by zero-padding the reference microphone representations with two time-frequency frames at the beginning and trimming the same length from the end, where the two-frame padding is due to the overlap-add process in reconstruction.

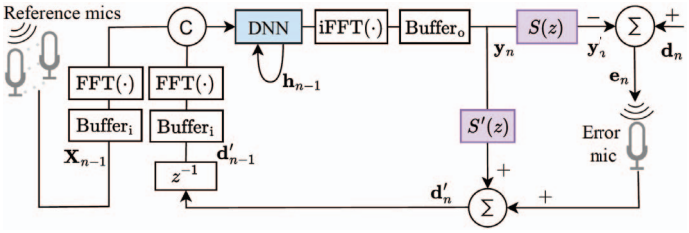

Fig. 1: DNN-based ANC performing frame-by-frame inference. $\mathbf{h}_{n-1}$ denotes internal representations of the DNN. Buffer$_i$ and Buffer$_o$ denote the input and output buffers for handling a long analysis window and the overlap-add reconstruction, respectively. HB concatenates spectral representations as input.

The DNN-based FF system models the primary path, and its difficulty escalates further when combined with the task of future prediction necessary for causality constraints [29]. This work includes the causal FF control structure as part of the unified framework.

3.2. DNN-based FB Control Structure

Wearables, such as headsets, can capture residual signals entering the ear when an error microphone is placed near the ear canal opening. When the loudspeaker is positioned close to the ear canal, it creates a low-latency secondary path. These factors enable the development of FB systems that rely only on error microphone signals. Similarly, the FB control module can be achieved by utilizing a DNN:

$$\mathbf{y}_n = f_{\text{FB}}(\mathbf{e}_{n-1}, \cdots, \mathbf{e}_1). \tag{3}$$

Rather than processing the error microphone signal sample by sample as in [20], [21], here we conduct frame-based processing, and the prediction scope can be explicitly controlled by the short synthesis window length discussed in Section 3.1. The FB algorithm regenerates the primary noise at the error microphone, denoted as $d'(t)$, and its frame-based representation is $\mathbf{d}'_n$. The regenerated noise signal is then used to predict the control signal for the subsequent frame. However, this recursive nature leads to a slow training process. This situation changes if we assume a perfectly estimated secondary path $S'(z)$, leading to a regenerated signal that perfectly matches the primary disturbance, $d'(t) = d(t)$. Thus, the FB system can be treated as sequence learning during training, akin to autoregressive modeling, while during inference, it can optionally operate in a streaming mode when $d'(t) \neq d(t)$, i.e., processing signals frame by frame as illustrated in Fig. 1.

While the adaptive FB system needs to consider stability and convergence behavior [6], [14], [30], the fixed-filter methods are alternative solutions. The DNN-based FB ANC operates similarly to pre-optimized control filters but with a much more powerful predictive ability powered by its data-driven nature.

3.3. DNN-based HB Control Structure

The HB systems combine the FF and FB control topologies. While it is possible to simply combine the two independent DNN modules, this would increase the system complexity and may require an additional merging module. Instead, we jointly process the signals

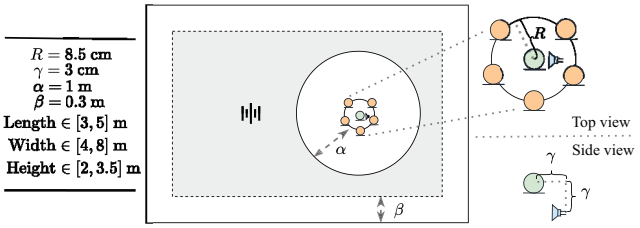

Fig. 2: Illustration of a simulation setup with 5 reference microphones, 1 error microphone, and 1 loudspeaker.

from the reference and error microphones and learn a DNN to output the spectral representations of a single control signal:

$$\mathbf{y}_n = f_{\text{HB}}(\mathbf{X}_{n-1},\cdots,\mathbf{X}_1;\mathbf{e}_{n-1},\cdots,\mathbf{e}_1). \quad (4)$$

The joint processing has the benefit of predicting with a broad understanding of the zone around the cancellation point. In contrast, the FF system needs to infer spatial characteristics at the error microphone based on the reference microphone signals, whereas the FB system has only the information of the cancellation point without spatial knowledge of the surrounding space.

Despite progress in DNN-based FF [16], [17] and FB [19] systems, independent development with inconsistent settings and limited evaluations leads to an incomplete understanding of their relative performance. Additionally, the DNN-based HB system remains unexplored. Neither its cancellation behavior nor its comparative performance against FB and FF have been thoroughly investigated. This work explores these DNN-based methods in a unified framework and presents a variety of evaluations to reveal the challenges involved.

4. EXPERIMENTS

Data Simulation: We simulate a uniform circular microphone array with a radius of R and a centered error microphone, resembling a user wearing smart glasses or headsets [31], [32]. The loudspeaker is placed close to the error microphone with a vertical and horizontal shift of γ. We synthesize disturbing sources using speech or noise databases from the deep noise suppression (DNS) challenge [33]. We randomly split the noise database into around 56k/7k/1.3k samples for training/validation/testing to train a noise-based ANC. This test set is called *DNS Noise*. We further evaluate the noise-based ANC model on a test set consisting of 300 samples randomly drawn from the NOISEX-92 dataset, including `engine`, `babble`, and `factory` [34] as in [16], [19], referred to as *NOISEX-92 Noise*. Furthermore, we use the speech database from the DNS challenge to train a speech-based ANC model, following the same data split for training/validation/testing. The speech test set is called *DNS Speech*. First, we create a *fixed setup* as in [16], [17], [19], where we simulate a room of dimension $3\times4\times2.5$ m using pyroomacoustics [35]. The disturbing sound is positioned at $1.5\times1.3\times1.0$ m, and the microphone array is centered at $1.5\times2.5\times1.0$ m. Beyond the fixed setup, we randomly sample the disturbing source position from the gray area and vary the room size, as illustrated in Fig. 2. The microphone array is then placed near the center of a room with a small deviation uniformly sampled from [-0.1, 0.1] m. Another important aspect is the simulation method to create room impulse responses (RIRs). The IS method [22] can accurately model early reflections by considering specular reflections. The RT method [23] can simulate diffuse reflection effects by generating sound rays in random directions to better simulate complex reflection patterns. This offers a more realistic simulation of late reflections. Therefore, unlike previous work using only the IS method [16], [17], [19], we use the hybrid simulator that combines the IS and RT methods to examine the impact of more realistic acoustic simulation on the performance of the broadband ANC problem. Unless explicitly stated otherwise, we use the hybrid simulator.

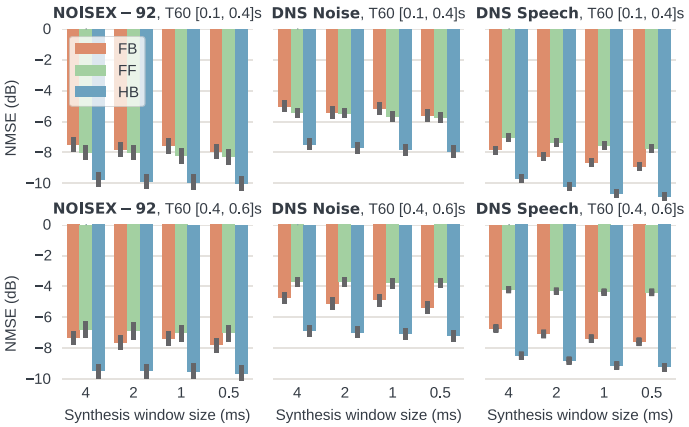

Fig. 3: Performance obtained on three test sets with a fixed source position and room size. Values are shown in mean and 95%-confidence interval (CI).

Experimental Setup: All data have a sampling rate of 16kHz. We transform data into their STFT representations, with an analysis window length of 20 ms. We use a short synthesis window as discussed in Section 3.1 and investigate four different synthesis window lengths: 4 ms, 2 ms, 1 ms, 0.5 ms. The hop size is set to half of the synthesis window length. We use asymmetric Hann windows for the dual-window STFT [36]–[38]. The analysis window combines the left and right halves of two square-root Hann windows, as detailed in [37]. Power normalization is used at the synthesis stage to ensure a perfect reconstruction of the STFT.

We employ the same UNet architecture as in [16] and adapt the feature dimension of the encoder's first convolutional layer for each control structure: FF/FB/HB uses 5/1/6-channel signals. The real and image parts of the STFT representations are stacked to form a feature dimension. All models are trained with a MSE loss applied to the anti-sound $y'(t)$ and $d(t)$ in the time domain. We train the DNN with the Adam optimizer [39]. We set an initial learning rate of 10^{-4} and half it every 25 epochs. In total, the DNN is trained for 100 epochs and the weights that achieve the best validation loss are saved for inference. We evaluate the performance using a widely used performance measure normalized MSE (NMSE): $10\log_{10}\frac{\sum_t(d(t)-y'(t))^2}{\sum_t d^2(t)}$. The lower the NMSE is, the better the model performs. When ANC is off, NMSE equals 0.

Results: We first evaluate the ANC systems (trained on the T60 range of [0.1,0.6]s) on the three test sets of the fixed setup in low ([0.1,0.4]s) and medium ([0.4,0.6]s) reverberation, as shown in Fig. 3. **(1)** Predicting short frames is highly effective for signals with predictable contents like voiced segments in speech (*DNS Speech*), but less so for highly nonstationary noise (*DNS Noise*). We observe that FF is basically unaffected by the synthesis window length, probably because the FF's performance is largely impacted by the modeling accuracy of the primary path. Last, for the relatively simple *NOISEX-92 Noise*, the models may easily predict 4 ms ahead, thus showing insignificant gains with a reduced synthesis window length. **(2)** FF and FB exhibit performance degradation when the reverberation increases, with the decrease being more pronounced on *DNS Noise* and *DNS Speech*, and less severe on the simpler *NOISEX-92 Noise*. FF and FB perform comparably on the three test sets in low reverberation, while FB is more robust than FF in medium reverberation. This is because FF needs to predict spatial characteristics at the location of the error microphone based on the primary noise captured by the reference microphones. While FF can perform this task reasonably well in environments with low reverberation, it struggles to perform robustly under higher reverberation

Table 1: Performance on NOISEX-92 (T60$\in[0.1,0.4]$ s) is shown as mean and 95 %-CI. * indicates a system failure (FB, HB) or not applicable (FF).

Distortions	Control Structures		
	FF	FB	HB
Matched secondary path	-8.22±0.34	-7.33±0.35	-9.83±0.34
Low pass (1k Hz)	-5.11±0.52/*	* / -5.03±0.52	* / -5.57±0.57
Low pass (500 Hz)	-4.63±0.48/*	* / -4.64±0.49	* / -4.96±0.53
Linear gain (±2 dB)	-6.86±0.25/*	-7.22±0.35 / -6.22±0.27	-9.42±0.32 / -7.95±0.22
Linear gain (±3 dB)	-5.68±0.20/*	-7.08±0.35 / -5.18±0.22	-8.95±0.31 / -6.44±0.18
Loudspeaker shift (1 cm)	-7.88±0.33/*	-7.31±0.35 / -6.99±0.35	-9.68±0.34 / -9.22±0.34
Loudspeaker shift (1.5 cm)	-7.06±0.34/*	* / -6.28±0.35	-9.25±0.36 / -8.13±0.38

conditions. This difficulty arises mainly because the late reverberation of RIRs, featuring stochastic scattering effects modeled by RT, complicates the accurate prediction of the control signal. This aspect will be explored further in Table 2. **(3)** We observe that HB outperforms FB and FF across all acoustic conditions. This may be primarily because the HB algorithm uses both the reference microphone recordings and the reconstructed signals at the error microphone, which implies a better understanding of how sound energy is distributed around the cancellation point. Consequently, it enables the model to handle early and late reflections more effectively. **(4)** Overall, future signal prediction remains a bottleneck for all approaches. The HB algorithm has a superior spatial understanding of the space and surpasses FF and FB, where the former struggles to accurately model the primary path, especially under increased reverberant conditions, while the latter lacks spatial understanding around the cancellation point and focuses only on spectral content prediction, as it relies only on signal channel signals. For the following experiments, we set the synthesis window length to 1 ms.

We investigate the models' robustness against distorted secondary paths [15] under the fixed setup. Table 1 presents the results as slash-separated values: the first represents a condition where the system continues using the previously perfectly estimated $\widehat{S}(z)$, even though $S(z)$ has changed in practice; the second applies the same distortion to both. The distorted secondary path may lead to $d'(t)\neq d(t)$. Thus, we perform frame-by-frame inference as illustrated in Fig. 1. We simulate three types of distortions: second-order low-pass filtering with cutoff frequencies of 500 Hz and 1 kHz; applying a multiplicative linear factor randomly sampled from the range $[\frac{1}{f},f]$, where f is derived from a gain of 2 and 3 dB; randomly moving the loudspeaker within the sphere of radius 1 and 1.5 cm around the original training position. **(1)** FF does not require $\widehat{S}(z)$ for signal regeneration but can be affected by distorted $S(z)$. For each type of distortion, the performance decrease of FF correlates with the severity, with more severe distortions causing greater degradation. However, severe mismatch of $S(z)$ (e.g., low-pass filtering and loudspeaker shifting) may cause system failures for FB and HB due to constructive interference of anti-noise and numerical instabilities of DNNs in their recursive loops. **(2)** When FB and HB systems do not experience system failures caused by distorted $S(z)$, they exhibit less performance decrease than FF, possibly because their recursive signal regeneration process compensates for some mismatches. **(3)** Applying the same distortion to $S(z)$ and $\widehat{S}(z)$ creates a mismatch with the training conditions. We observe that the FB and HB systems are still able to handle mild mismatches, albeit with some degradation in performance. **(4)** HB maintains the advantages over FB for most distortions, because the primary noise captured by the reference microphones remains unaffected by the distorted secondary paths.

Finally, we move away from the *fixed setup* with the *hybrid simulator*, which is shown in rows *i*, *I*, and *a* of Table 2 as reference, and investigate how the fully DNN-based control structures learn under various acoustic conditions, assuming undistorted and perfectly estimated secondary paths. We progressively replace the more practical hybrid simulator that combines the IS and RT methods with the IS method only (*IS only*=True), alter the source position (*Fix SP*=False),

Table 2: Performance of various setups in the T60 range $[0.2,0.4]$s (mean and 95%-CI). *RS*: room size; *SP*: source position; *IS*: image-source method.

		Simulation setup			Noise	
		Fix RS	Fix SP	IS only	DNS Noise	NOISEX-92
FF	*i*	✓	✓	✗	-5.61±0.17	-8.62±0.35
	ii	✓	✓	✓	-11.07±0.23	-15.58±0.41
	iii	✓	✗	✓	-4.86±0.27	-11.37±0.61
	iv	✗	✗	✓	-3.53±0.22	-7.13±0.44
	v	✗	✗	✗	-2.50±0.17	-5.07±0.38
FB	*I*	✓	✓	✗	-5.74±0.25	-8.32±0.32
	II	✓	✓	✓	-6.33±0.27	-10.94±0.40
	III	✓	✗	✓	-5.52±0.28	-11.26±0.43
	IV	✗	✗	✓	-4.79±0.25	-8.15±0.35
	V	✗	✗	✗	-4.48±0.24	-6.34±0.30
HB	*a*	✓	✓	✗	-8.08±0.22	-10.49±0.32
	b	✓	✓	✓	-11.23±0.23	-15.72±0.41
	c	✓	✗	✓	-6.39±0.28	-12.90±0.49
	d	✗	✗	✓	-5.01±0.25	-8.51±0.36
	e	✗	✗	✗	-4.64±0.24	-6.59±0.31

and vary the room size (*Fix RS*=False). **(1)** When disregarding the stochastic effects of late reverberations modeled by the RT method and simulating the acoustic environments by only the IS method, its deterministic nature is relatively easier for DNNs to learn, thus providing higher gains for all ANC models when training and testing in this setup (see rows *i* and *ii*; *I* and *II*; *a* and *b*). As FB uses only temporal-spectral information and is mainly constrained by signal predictability, only a marginal performance gain is observed on highly non-stationary *DNS Noise* (*I* and *II*). In contrast, this acoustic simplification favors FF significantly. As specular reflections of rooms are relatively easier for FF to model, it can better predict the spatial characteristics at the error microphone. Thus, we observe that training and testing under this condition results in considerably high performance gains (*i* and *ii*). A similar trend can be observed for HB (*a* and *b*). Additionally, HB tends to degenerate into FF due to the dominant performance of FF in this simplified condition. **(2)** It becomes more challenging for FF when the source position is not fixed (*iii*). In contrast, HB remains relatively more robust (*iii* and *c*), which is due to less performance variation exhibited by FB under varying acoustic conditions (*III*). **(3)** Further performance decrease of FF can be observed when the room size and source position vary simultaneously (*iv*) and meanwhile switch back to the hybrid simulator (*v*). This highlights that FF is susceptible to environmental changes, as it relies on both accurate spatial modeling and room geometry information. Consequently, HB degenerates into FB (*d* and *IV*; *e* and *V*) when FF struggles to model the complex primary paths. **(4)** Overall, it is crucial to explicitly incorporate the spatial information of the disturbances and the room geometry, while also modeling the stochastic nature of reverberant RIR tails, to develop a robust open-ear ANC system.

5. CONCLUSION

The goal of this paper is to provide a broad understanding of the effectiveness of DNN-based ANC under various acoustic conditions. For this, we have presented a unified framework for a systematic evaluation of the frame-based FF and FB systems and proposed a simple DNN-based HB extension. We have shown that FF performs reasonably well when the primary path and acoustic environment are modeled accurately, while FB maintains more consistent performance in various settings. We observe the challenge of handling stochasticity effects in reverberant conditions, especially for FF, and that future prediction is a bottleneck for both systems, especially when dealing with highly non-stationary disturbances. While the recursive loop of FB can handle mild secondary path distortions, severe mismatches may lead to system failures. Overall, the proposed HB extension exhibits a more robust performance. It can incorporate the benefits of both FF and FB when acoustic modeling is accurate, while still retaining the performance of FB when the primary path is difficult to capture.

REFERENCES

[1] P. Nelson, A. Curtis, S. Elliott, and A. Bullmore, "The active minimization of harmonic enclosed sound fields, part i: Theory," *Journal of Sound and Vibration*, vol. 117, no. 1, pp. 1–13, 1987.

[2] S. M. Kuo and D. R. Morgan, "Active noise control: A tutorial review," *Proceedings of the IEEE*, vol. 87, no. 6, pp. 943–973, 1999.

[3] S. M. Kuo and D. Morgan, *Active Noise Control Systems: Algorithms and DSP Implementations*. John Wiley & Sons, Inc., 1995.

[4] Y. Kajikawa, W.-S. Gan, and S. M. Kuo, "Recent advances on active noise control: Open issues and innovative applications," *APSIPA Transactions on Signal and Information Processing*, vol. 1, p. e3, 2012.

[5] S. Liebich, J.-G. Richter, J. Fabry, C. Durand, J. Fels, and P. Jax, "Direction-of-arrival dependency of active noise cancellation headphones," in *Noise Control and Acoustics Division Conference*, vol. ASME 2018 Noise Control and Acoustics Division Session presented at INTERNOISE 2018, 2018, p. V001T08A003.

[6] S. Liebich, J. Fabry, P. Jax, and P. Vary, "Signal processing challenges for active noise cancellation headphones," in *Speech Communication; 13th ITG-Symposium*, 2018, pp. 1–5.

[7] P. R. Benois, P. Nowak, and U. Zoelzer, "Hybrid active noise control structures: A short overview," in *Speech Communication; 13th ITG-Symposium*, 2018, pp. 1–5.

[8] O. J. Tobias and R. Seara, "Performance comparison of the FxLMS, nonlinear FxLMS and leaky FxLMS algorithms in nonlinear active control applications," in *European Signal Proc. Conf. (EUSIPCO)*, 2002, pp. 1–4.

[9] B. Xu and S. D. Sommerfeldt, "Generalized acoustic energy density based active noise control in single frequency diffuse sound fields," *J. Acoust. Soc. Am.*, vol. 136, no. 3, pp. 1112–1119, 2014.

[10] D. Shi, W.-S. Gan, B. Lam, and S. Wen, "Feedforward selective fixed-filter active noise control: Algorithm and implementation," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 28, pp. 1479–1492, 2020.

[11] D. Shi, B. Lam, K. Ooi, X. Shen, and W.-S. Gan, "Selective fixed-filter active noise control based on convolutional neural network," *Signal Processing*, vol. 190, p. 108317, 2022.

[12] Z. Luo, D. Shi, and W.-S. Gan, "A hybrid SFANC-FxNLMS algorithm for active noise control based on deep learning," *IEEE Signal Proc. Letters*, vol. 29, pp. 1102–1106, 2022.

[13] D. Chen, L. Cheng, D. Yao, J. Li, and Y. Yan, "A secondary path-decoupled active noise control algorithm based on deep learning," *IEEE Signal Proc. Letters*, vol. 29, pp. 234–238, 2021.

[14] S. Snyder and C. Hansen, "The effect of transfer function estimation errors on the filtered-x LMS algorithm," *IEEE Transactions on Signal Processing*, vol. 42, no. 4, pp. 950–953, 1994.

[15] Y. Song, Y. Gong, and S. M. Kuo, "A robust hybrid feedback active noise cancellation headset," *IEEE Trans. on Speech and Audio Proc.*, vol. 13, no. 4, pp. 607–617, 2005.

[16] H. Zhang and D. Wang, "Deep ANC: A deep learning approach to active noise control," *Neural Networks*, vol. 141, pp. 1–10, 2021.

[17] H. Zhang, A. Pandey *et al.*, "Low-latency active noise control using attentive recurrent network," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 31, pp. 1114–1123, 2023.

[18] H. Zhang and D. Wang, "Deep MCANC: A deep learning approach to multi-channel active noise control," *Neural Networks*, vol. 158, pp. 318–327, 2023.

[19] Y.-J. Cha, A. Mostafavi, and S. S. Benipal, "DNoiseNet: Deep learning-based feedback active noise control in various noisy environments," *Engineering Applications of Artificial Intelligence*, vol. 121, p. 105971, 2023.

[20] A. Bayestehtashk, A. Kumar, and M. Wurtz, "Design of feedback active noise cancellation filter using nested recurrent neural networks," in *Proc. Interspeech*, 2024, pp. 3270–3274.

[21] D. Wu, X. Wu, and T. Qu, "A hybrid deep-online learning based method for active noise control in wave domain," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2024, pp. 1301–1305.

[22] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.

[23] D. Schröder, *Physically based real-time auralization of interactive virtual environments*. Logos Verlag Berlin GmbH, 2011, vol. 11.

[24] F. Yang, Y. Cao, M. Wu, F. Albu, and J. Yang, "Frequency-domain filtered-x LMS algorithms for active noise control: A review and new insights," *Applied Sciences*, vol. 8, no. 11, p. 2313, 2018.

[25] Y. Iotov, S. M. Nørholm, V. Belyi, M. Dyrholm, and M. G. Christensen, "Computationally efficient fixed-filter ANC for speech based on long-term prediction for headphone applications," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2022, pp. 761–765.

[26] Y. Kajikawa and R. Hirayama, "Feedback active noise control system combining linear prediction filter," in *European Signal Proc. Conf. (EUSIPCO)*, 2010, pp. 31–35.

[27] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 26, no. 10, pp. 1702–1726, 2018.

[28] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and T. Sainath, "Deep learning for audio signal processing," *IEEE J. of Selected Topics in Signal Proc.*, vol. 13, no. 2, pp. 206–219, 2019.

[29] L. Zhang and X. Qiu, "Causality study on a feedforward active noise control headset with different noise coming directions in free field," *Applied Acoustics*, vol. 80, pp. 36–44, 2014.

[30] T. Schumacher, H. Krüger, M. Jeub, P. Vary, and C. Beaugeant, "Active noise control in headsets: A new approach for broadband feedback ANC," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2011, pp. 417–420.

[31] L. Alzubaidi, J. Bai, A. Al-Sabaawi, J. Santamaría, A. S. Albahri, B. S. N. Al-Dabbagh, M. A. Fadhel, M. Manoufali, J. Zhang, A. H. Al-Timemy *et al.*, "A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications," *Journal of Big Data*, vol. 10, no. 1, p. 46, 2023.

[32] J. Donley, V. Tourbabin, J.-S. Lee, M. Broyles, H. Jiang, J. Shen, M. Pantic, V. K. Ithapu, and R. Mehra, "Easycom: An augmented reality dataset to support algorithms for easy communication in noisy environments," *arXiv preprint*, 2021. [Online]. Available: https://arxiv.org/abs/2107.04174

[33] C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun, P. Rana, S. Srinivasan, and J. Gehrke, "The INTERSPEECH 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," in *Proc. Interspeech*, 2020, pp. 2492–2496.

[34] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[35] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2018, pp. 351–355.

[36] D. Mauler and R. Martin, "A low delay, variable resolution, perfect reconstruction spectral analysis-synthesis system for speech enhancement," in *European Signal Proc. Conf. (EUSIPCO)*, 2007, pp. 222–226.

[37] S. U. Wood and J. Rouat, "Unsupervised low latency speech enhancement with RT-GCC-NMF," *IEEE J. of Selected Topics in Signal Proc.*, vol. 13, no. 2, pp. 332–346, 2019.

[38] Z.-Q. Wang, G. Wichern, S. Watanabe, and J. Le Roux, "STFT-domain neural speech enhancement with very low algorithmic latency," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 31, pp. 397–410, 2022.

[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Int. Conf. Learning Repr. (ICLR)*, 2015.