Low-Latency Active Noise Control Using Attentive Recurrent Network

Hao Zhang, *Member, IEEE*, Ashutosh Pandey, *Member, IEEE*, and De Liang Wang, *Fellow, IEEE*

***Abstract*—Processing latency is a critical issue for active noise control (ANC) due to the causality constraint of ANC systems. This paper addresses low-latency ANC in the context of deep learning (i.e. deep ANC). A time-domain method using an attentive recurrent network (ARN) is employed to perform deep ANC with smaller frame sizes, thus reducing algorithmic latency of deep ANC. In addition, we introduce a delay-compensated training to perform ANC using predicted noise for several milliseconds. Moreover, a revised overlap-add method is utilized during signal resynthesis to avoid the latency introduced due to overlaps between neighboring time frames. Experimental results show the effectiveness of the proposed strategies for achieving low-latency deep ANC. Combining the proposed strategies is capable of yielding zero, even negative, algorithmic latency without affecting ANC performance much, thus alleviating the causality constraint in ANC design.**

***Index Terms*—Active noise control, deep ANC, algorithmic latency, ARN, low-latency.**

I. Introduction

NOISE is an auditory annoyance that has negative effects on human listeners and is recognized as a type of pollution. Two different strategies exist for controlling noise: passive and active noise control. Passive noise control is the traditional way to reduce noise, and it achieves noise attenuation using passive methods like insulation and silencers. Active noise control (ANC) is a noise cancellation technology based on the principle of destructive superposition of acoustic signals. It works by generating an anti-noise with the equal amplitude and opposite phase of the primary (unwanted) noise, hence resulting in the cancellation of both when they are superposed at an error microphone [1]. ANC has attracted increasing attention in research over the past few decades and has been used in automobiles [2], headphones [3], airplanes [4], and so on [5], [6].

Conventionally, ANC is accomplished by optimizing controller weights using adaptive filters so that the error signal is minimized [7]. Filtered-x least mean square (FxLMS) and its extensions are the most widely used ANC algorithms. They work by estimating a secondary path beforehand and then filtering the reference noise with the estimated secondary path before feeding it to the controller [8]. However, nonlinear distortions are inevitably introduced due to the use of electronic devices like loudspeakers [9], [10]. The adaptive filter approach is fundamentally linear and does not perform satisfactorily in the presence of nonlinear distortions [11].

Recently, deep learning has been utilized for fixed-parameter ANC [12] considering the capacity of deep neural networks in modeling complex nonlinear relationships [13], [14], [15], [16], [17], [18]. In a previous study, we formulated ANC as a supervised learning problem for the first time and proposed a deep learning approach, called deep ANC, to address the nonlinear ANC problem [13], [14]. Subsequently, a deep learning based selective fixed-filter ANC method that employs a convolutional neural network for noise type classification and control filter selection was introduced in [16]. Later, Chen et al. proposed a secondary path-decoupled ANC method (SPD-ANC) using two pre-trained convolutional recurrent networks to address the nonlinearity of the secondary path [17]. More recently, we expanded the single-channel deep ANC to the multi-channel domain and developed a deep learning approach for active noise control at multiple spatial points and within a spatial zone [15]. All these deep learning based methods can be viewed as fixed-parameter ANC and they achieve active noise control by training a deep neural network (DNN) offline. Compared to traditional fixed-parameter ANC methods, deep ANC is capable of nonlinear active noise reduction for a variety of noises through large-scale multi-condition training.

A unique constraint of ANC is that it targets noise in physical space unlike, say, noise reduction in mobile communication. Specifically, the error microphone of an ANC system adds primary noise and anti-noise signals arriving at its location acoustically. This leads to the *causality constraint* of ANC systems; that is, the sum of controller processing time and the secondary path acoustic delay must be no greater than the primary path acoustic delay (the time for noise to propagate along the primary path) [19], [20]. Many studies have demonstrated the effects of causality on the performance of ANC systems. Burdisso et al. investigated system causality and developed a formulation to carry out causality analysis of feedforward ANC systems subjected to broadband excitations [21]. Kong and Kuo studied the efficiency of ANC systems for ducts under non-causal conditions [19]. Zhang and Qiu presented a causality study on a typical feedforward ANC headset and systematically analyzed

Manuscript received 27 June 2022; revised 24 December 2022 and 30 January 2023; accepted 1 February 2023. Date of publication 13 February 2023; date of current version 27 February 2023. This work was supported by NIH\NIDCD under Grant R01 DC012048 06-10. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Romain Serizel. *(Corresponding author: Hao Zhang.)*

Hao Zhang and Ashutosh Pandey are with the Department of Computer Science and Engineering, Ohio State University, Columbus, OH 43210-1277 USA (e-mail: zhang.6720@osu.edu; pandey.99@osu.edu).

De Liang Wang is with the Department of Computer Science and Engineering and the Center for Cognitive and Brain Sciences, Ohio State University, Columbus, OH 43210-1277 USA (e-mail: dwang@cse.ohio-state.edu).

Digital Object Identifier 10.1109/TASLP.2023.3244528

the performance of ANC headsets in terms of delays [20]. Kurczyk and Pawelczyk addressed the latency problem of ANC systems by using soft computing algorithms, including fuzzy inference. [22], [23], [24]. Effects of primary source locations and microphone locations on causal configuration and ANC performance are studied in [25], [26].

The causality constraint is a dominant factor in the design of ANC systems, and it must be satisfied to perform noise attenuation. However, block-based algorithms such as frequency-domain FxLMS and deep ANC, possess an algorithmic delay determined by the frame size since they are implemented in a frame-by-frame manner [27], [28]. This delay could violate the causality constraint and is considered a major limitation for block-based ANC algorithms. The connection between time-domain and frequency-domain effort weighting in ANC design was introduced in [29]. Yang et al. studied the delays introduced by frequency-domain ANC methods and presented different schemes for addressing these delays [30]. Shi et al. utilized a virtual-sensing technique for frequency-domain multi-channel ANC to satisfy the causality constraint between the locations of the physical and virtual microphones [31]. A training strategy using predicted anti-noise to compensate for the delay of the frequency-domain ANC method was proposed in [14]. Although many studies have been proposed for latency reduction, the latency problem remains for block-based ANC algorithms.

Time-domain methods have been recently proposed for supervised speech separation. Compared to frequency-domain methods that use time-frequency representations for extracting input features and training targets, time-domain methods directly predict target signal samples from input signal samples, and can enhance magnitude and phase jointly in the process [32], [33]. In addition, frequency-domain methods usually require a relatively longer frame size to ensure an adequate frequency resolution, leading to longer algorithmic latencies. There is no such limitation for time-domain methods and they can be implemented using smaller frame sizes. Fu et al. proposed a time-domain network to optimize the short-term objective intelligibility metric [34]. A fully convolutional time-domain audio separation network was introduced in [35] for end-to-end time-domain speaker separation, and the frame size can be set as small as 2 ms. A convolutional neural network with a frequency-domain loss was proposed in [36] to address speech enhancement in the time domain. Very recently, Pandey and Wang proposed an attentive recurrent network (ARN) for time-domain speech enhancement [37]. Time-domain methods are potentially more suitable for achieving low-latency deep ANC.

Building on deep ANC, this paper aims at achieving low latencies by reducing the algorithmic latency of deep ANC. The contributions of this paper are summarized below. First, we introduce time-domain deep ANC utilizing an attentive recurrent network [37], which enables the implementation of deep ANC with smaller frame sizes. Second, to counter algorithmic latency, a delay-compensated training strategy is proposed to perform ANC using noise predicted ahead of time. Third, a revised overlap-add (OLA) method is utilized during signal resynthesis to avoid the latency introduced by overlaps between neighboring frames. Finally, we combine the proposed strategies to achieve deep ANC with zero or even negative algorithmic latency. The proposed approach represents a big stride towards alleviating the causality constraint of ANC and expanding the scope of ANC design.

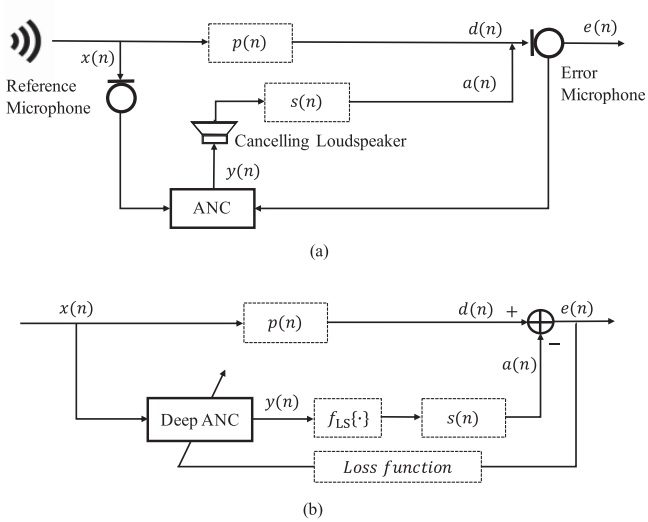

Fig. 1. Diagrams of (a) feedforward ANC system, and (b) deep ANC approach.

A preliminary version of this study has recently been accepted for conference presentation [38]. Compared to the conference version, this paper conducts more extensive evaluations, investigates different strategies, and provides insights into combining the proposed strategies for low-latency deep ANC.

The remainder of this paper is organized as follows. Section II introduces the signal model and deep ANC method. Section III describes the proposed low-latency deep ANC techniques. The experimental setup is given in Section IV. Section V provides the evaluation results and comparisons. Section VI concludes the paper.

II. Active Noise Control

A. Signal Model

A typical feedforward ANC system consists of a reference microphone, a canceling loudspeaker, and an error microphone, as shown in Fig. 1(a). The primary noise picked up by the error microphone, $d(n)$, is generated by convolving the reference noise with the primary path. The reference signal $x(n)$ sensed by a reference microphone is fed to the active noise controller to generate a canceling signal $y(n)$. The canceling signal is then played by a canceling loudspeaker and propagated through the secondary path to get an anti-noise, $a(n)$, which then cancels or attenuates the primary noise $d(n)$. The corresponding error signal received at the error microphone is obtained as

$$\begin{aligned} e(n) &= d(n) - a(n) \\ &= p(n) * x(n) - s(n) * f_{\text{LS}}\{y(n)\} \end{aligned} \tag{1}$$

where n is the time index, $p(n)$ and $s(n)$ denote the primary and secondary path, respectively, $f_{\text{LS}}\{\cdot\}$ denotes the function of

a loudspeaker, and symbol $*$ denotes convolution. Note that the anti-noise is subtracted in (1) to achieve noise cancellation.

B. Causality Constraint of a Feedforward ANC System

To achieve noise attenuation, the anti-noise has to reach the error microphone no later than the primary noise. In other words, the total delay of the controller and the secondary path should not be larger than that of the primary path. This is the so-called causality constraint, and it must be satisfied or the primary noise cannot be reduced by the system.

The causality constraint can be expressed as

$$T_p \geq T_{\text{ANC}} + T_s \;\text{ or }\; T_{\text{ANC}} \leq T_p - T_s \tag{2}$$

where T_p and T_s denote the acoustic delays introduced by the primary and secondary paths, respectively, which are proportional to the lengths of the corresponding paths. T_{ANC} denotes the latency introduced by the controller, which equals the sum of the ANC processing latency and the total system delay (including those of A/D and D/A converters, and loudspeaker) [19].

The causality constraint given in (2) is obtained from the fundamental wave propagation point of view. The practical constraint of an ANC system also depends on noise type and prediction. For example, a tonal noise is easy to predict, and causality would not be an issue in this case. The system causality is also affected by the ANC configuration as well as the processing latency of ANC algorithms. Positions of loudspeaker and microphones determine the maximum latency allowed for T_{ANC}, and they need to be chosen carefully in the design of an ANC system. With a given ANC configuration, the algorithmic delay of ANC should be controlled to be as low as possible to guarantee the system causality.

C. Deep ANC

Unlike traditional ANC methods, which require estimating the secondary path and the adaptive filter individually, deep ANC trains a DNN using large-scale multi-condition training to directly approximate an optimal controller that minimizes the error signal in a variety of noisy environments [14]. A diagram of the deep ANC approach is given in Fig. 1(b). The overall goal is to estimate a canceling signal from the reference signal so that the corresponding anti-noise cancels the primary noise. Deep ANC takes as input a reference signal and sets the ideal anti-noise as the training target. The ideal anti-noise is the same as the primary noise to accomplish complete noise cancellation. During training, the output of deep ANC is treated as an "intermediate product," and the anti-noise is produced by passing the output through the loudspeaker and secondary path. The loss function calculated from the error signal is then used to guide model training.

III. Low-Latency Deep ANC

A. Algorithmic Latency of Deep ANC

Deep ANC is block-based, where signals are processed in a frame-by-frame manner. Specifically, an input signal is first chunked into short overlapping blocks of waveform samples and the blocks are then transformed into a sequence of frames. Taking a frequency domain method for example, each block in the input sequence is multiplied by an analysis window and then converted to the frequency domain using discrete Fourier transform (DFT). Resynthesis of a time-domain signal is achieved by taking the inverse DFT of the transformed frames, multiplying the obtained samples with a synthesis window, and combining neighboring frames using the OLA method [39]. These steps incur an algorithmic delay determined by the frame length and frame shift.

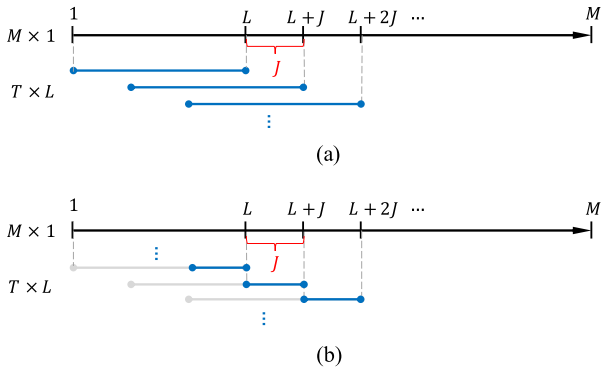

Fig. 2. Illustrations of (a) OLA, and (b) revised OLA. The corresponding algorithmic latency are L and J, respectively.

An illustration of OLA is given in Fig. 2(a), where a signal with M samples is chunked into T frames with a frame size of L and frame shift of J. Due to overlaps between neighboring frames, to fully synthesize a single sample, all frames that contain this sample need to be processed. For example, we have to wait till the end of the current frame to generate its initial $J-1$ samples, which results in a delay in the range of $(L-J, L]$ samples. In this paper, we define the algorithmic latency as the maximum delay, L, for simplicity.

B. ARN Based Time-Domain ANC

The most straightforward way of reducing algorithmic latency of deep ANC is to shorten the frame size. For frequency-domain methods, using a smaller frame size results in a lower frequency resolution and may degrade system performance [40]. We propose to realize deep ANC using ARN in the time domain, which can be easily implemented with smaller frame sizes. Further, we find ARN to be highly effective for the deep ANC task even with smaller frame sizes.

ARN is recently proposed in [41] for effectively incorporating an attention mechanism [42] into recurrent neural networks (RNNs). The processing flow of ARN based time-domain deep ANC is shown in Fig. 3(a). A reference noise $x(n)$ with M samples is first divided into T overlapping frames with a frame size of L and frame shift of J. Subsequently, a linear layer is used to project these frames to a representation of size N, which is then processed by a four-layered ARN. The output of the ARN is projected back to size L using another linear layer. Finally, OLA is utilized to obtain the waveform of canceling signal $y(n)$.

The architecture of an ARN layer used in this study is shown in Fig. 3(b). It comprises a recurrent neural network (RNN)

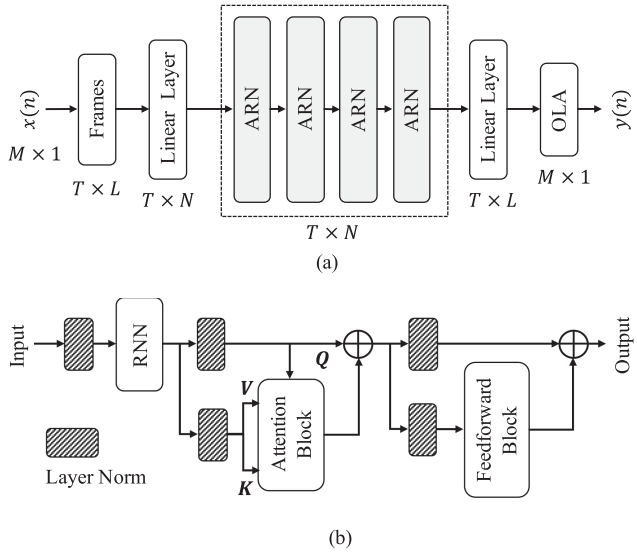

Fig. 3. Diagrams of (a) ARN based time-domain ANC, and (b) ARN architecture.

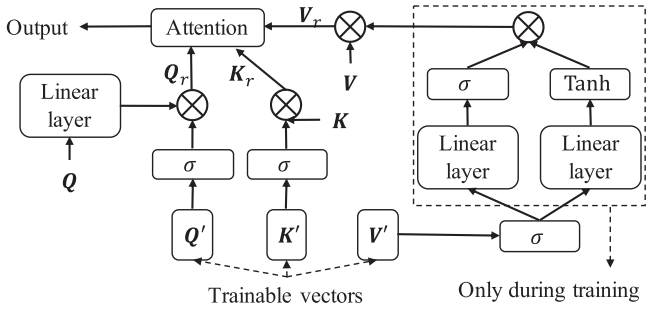

Fig. 4. Attention block in ARN.

with long short-term memory (LSTM), a self-attention block, and a feedforward block. The input to ARN is firstly layer normalized [43] and fed to an RNN. The output of the RNN is then normalized using two parallel layer normalizations, where the first layer normalized output is used as query ($\mathbf{Q}$), and the second one is used as key ($\mathbf{K}$) and value ($\mathbf{V}$) for the following attention block. The output of the attention block is added to $\mathbf{Q}$ to form a residual connection. Afterwards, the final output is normalized using two separate layer normalizations, in which one of the outputs is processed using the feedforward block and the other one is added to the output of the feedforward block in a residual way.

The attention block in ARN, shown in Fig. 4, takes $\{\mathbf{Q}, \mathbf{K}, \mathbf{V}\} \in \mathbb{R}^{T\times N}$ as inputs and comprises three trainable vectors $\{\mathbf{Q}', \mathbf{K}', \mathbf{V}'\} \in \mathbb{R}^{1\times N}$. A gating mechanism is utilized to refine the inputs as

$$\begin{aligned}\mathbf{K}_r &= \mathbf{K} \otimes \sigma\left(\mathbf{K}'\right)\\ \mathbf{Q}_r &= \mathrm{Lin}(\mathbf{Q}) \otimes \sigma\left(\mathbf{Q}'\right)\\ \mathbf{V}_r &= \mathbf{V} \otimes \left[\sigma\left(\mathrm{Lin}\left(\mathbf{V}'\right)\right) \otimes \mathrm{Tanh}\left(\mathrm{Lin}\left(\mathbf{V}'\right)\right)\right] \end{aligned} \tag{3}$$

where σ is sigmoidal nonlinearity, $\mathrm{Lin}()$ is a linear layer, and $\otimes$ denotes element-wise multiplication. Note that $\sigma(\mathrm{Lin}(\mathbf{V}')) \otimes \mathrm{Tanh}(\mathrm{Lin}(\mathbf{V}'))$ represents a constant vector computed from $\mathbf{V}$, and this operation is used during training for better optimization of $\mathbf{V}$. Once the model is trained, its value from the best model is used during evaluation.

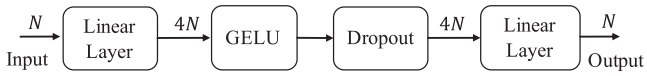

Fig. 5. Feedforward block in ARN.

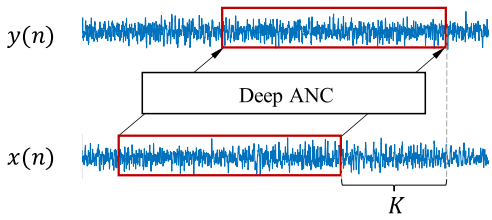

Fig. 6. Illustration of using deep ANC to predict K samples in advance.

The output of the attention block is obtained as

$$\mathbf{A} = \mathrm{Softmax}\left(\frac{\mathbf{Q}_r\mathbf{A}_r^T}{\sqrt{N}}\right)\mathbf{V}_r \tag{4}$$

The feedforward block in ARN is a fully connected network with one hidden layer of size $4N$, Gaussian error linear unit (GELU) nonlinearity [44], and dropout. A diagram of the block is shown in Fig. 5. A detailed description of the ARN can be found in [37].

C. Delay-Compensated Training

Another strategy for reducing latency is to perform ANC using predicted noise, and the resulting strategy is called delay-compensated training [14]. The main idea is to train a deep ANC model to predict the canceling signal a few samples ahead of time, thus compensating for the overall delay. During model training, instead of correctly aligning input and training target, we train the model to predict the target signal K samples in advance, as shown in Fig. 6. By prediction, the proposed strategy can cancel primary noise with K/f_s ms in advance, where fs denotes the sampling frequency. The delay-compensated training technique proposed in this paper extends the one introduced in [14], which predicts noise at the frame level (not the sample level in the present study).

A predicted noise is an estimate of the actual noise, and ANC using the predicted noise will lead to reduced performance compared to using the actual noise. However, the controller may be required to predict the primary noise if the causality constraint is violated. Such a strategy is useful for ANC tasks since it can intrinsically alleviate the causality constraint of ANC systems.

D. Revised Overlap-Add for Signal Resynthesis

Part of the algorithmic latency of block-based methods originates from the overlap-add procedure. Having overlaps between neighboring frames benefits from the averaging/pooling of multiple frames and results in a smoother estimate. We propose to revise the OLA method by setting a part of or the entire

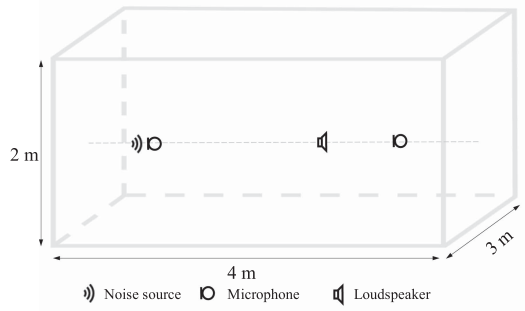

Fig. 7. Illustration of the ANC experimental setup.

overlapping samples to zero during signal resynthesis, in order to reduce the latency introduced by overlaps between neighboring frames. The revised OLA with all the overlapping samples set to zero, as shown in Fig. 2(b), reduces the algorithmic latency from frame size L to frame shift J. Considering the power of deep learning in prediction and that noise signals are relatively stationary and hence predictable, deep ANC with revised OLA has the potential to achieve low algorithmic latency without sacrificing ANC performance by much.

IV. Experimental Setup

A. Experimental Setting

Deep ANC is trained utilizing large-scale multi-condition training, exposing the ANC model to a large variety of noisy environments. To achieve a noise-independent model, we create a training set using 10000 non-speech environmental sounds from a sound-effect library (http://www.sound-ideas.com) [45]. Babble noise, engine noise, speech-shaped noise (denoted as "SSN"), and factory noise from NOISEX-92 dataset [46] are used for testing. The test noises are unseen during training, and hence can evaluate the generalization ability of the proposed method.

Many studies evaluate ANC systems for noise canceling in an enclosure [47], [48]. We follow the setup given in [14] and simulate a rectangular room of size 3 m × 4 m × 2 m (width × length × height) to carry out experiments. The primary and secondary paths are simulated as room impulse responses (RIRs) using the image method [49]. The reference microphone is located at the position (1.5, 1, 1) m, the canceling loudspeaker at (1.5, 2.5, 1) m, and the error microphone at (1.5, 3, 1) m. This experimental setup is illustrated in Fig. 7. Five reverberation times (T60 s) 0.15 s, 0.175 s, 0.2 s, 0.225 s, 0.25 s are used for generating training RIRs. The RIRs with reverberation time 0.2 s are used for testing. Their corresponding frequency responses are plotted in Fig. 8.

Nonlinear saturation effects are a common type of loudspeaker nonlinearity, and they can be simulated using the scaled error function (SEF) [10], [50]

$$f_{\text{SEF}}(y) = \int_0^y e^{-\frac{z^2}{2\eta^2}} dz, \tag{5}$$

where y is the input to the loudspeaker, and η^2 defines the strength of nonlinearity. The SEF becomes linear as η^2 tends to infinity, and a hard limiter as it tends to zero. In our experiments, the loudspeaker function $f_{\text{LS}}\{\cdot\}$ in (1) is implemented using $f_{\text{SEF}}(y)$. Our model is trained using four loudspeaker functions: $\eta^2 = 0.1$ (severe nonlinearity), $\eta^2 = 1$ (moderate nonlinearity), $\eta^2 = 10$ (soft nonlinearity), and $\eta^2 = \infty$ (linear). During training, we randomly select a loudspeaker function for each input signal, and generate the loudspeaker signal by passing a canceling signal through the loudspeaker function. For testing, both trained and untrained ($\eta^2 = 0.5$) loudspeaker functions are used.

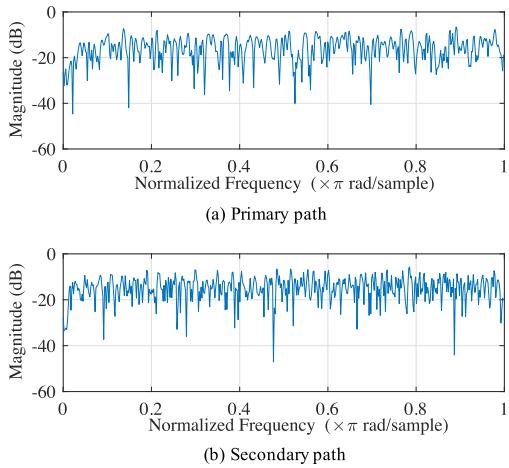

Fig. 8. Room frequency response with T60 = 0.2 s for (a) primary path, and (b) secondary path.

We create 20000 training signals and 100 test signals for each test noise. Each training noise is created by randomly cutting a 3-second segment from the concatenated signal of the 10000 non-speech sounds. The test noises are generated similarly from the 4 test noises. The primary noise at the error microphone is generated by convolving a source noise with a randomly selected RIR for the primary path. An estimated anti-noise is generated by passing the canceling signal successively through a loudspeaker function and the secondary path (see Fig. 1(b)). All the signals are resampled to 16 kHz.

Parameter N in ARN is set to 512, and a dropout rate of 5% is used in the feedforward block. Utterance level mean-squared error loss in the time domain is used for model training. The ARN model is trained using the Adam optimizer [51] with a learning rate of 0.0001 for 30 epochs.

B. Comparison Methods

We compare the proposed time-domain deep ANC method with FxLMS, tangential hyperbolic function based FxLMS (THF-FxLMS), [50], an optimal FxLMS solution [52], SPD-ANC [17], and a convolutional recurrent network (CRN) based frequency-domain method [14] in both linear and nonlinear situations.

FxLMS works by estimating a secondary path first and then applying it to the reference signal to compensate for the effect of the secondary path. It is a popular ANC algorithm due to its robust performance and ease of implementation. However, it

fails to identify the secondary path accurately in the presence of nonlinear distortions and consequently degrades the overall ANC performance. THF-FxLMS uses the tangent hyperbolic function (THF) to model the saturation effect of loudspeaker and then design the nonlinear ANC controller utilizing the predicted degree of nonlinearity [50]. It has been shown by Ghasemi et al. [50] that THF-FxLMS outperforms FxLMS for noise attenuation in situations with nonlinear distortions. In the optimal FxLMS solution [52], the ground-truth secondary path (including the distortions introduced by a canceling loudspeaker with nonlinearity) is used in ANC controller update and the steady-state results are presented for comparison.

The step sizes of FxLMS related methods in our experiments are chosen carefully for different noises according to the criteria given in [53] and [54] to ensure stable updating and good noise attenuation. Specifically, the step size used for updating FxLMS for babble noise, engine noise, SSN, and factory noise is set to 0.3, 0.05, 0.4, 0.4, respectively. The step size for updating THF-FxLMS is set to 0.3, 0.05, 0.4, 0.4, and for obtaining the optimal FxLMS solution is set to 0.25, 0.05, 0.2, 0.4 for the four noises, respectively. The algorithmic delay of all the adaptive ANC methods in the time domain is one signal sample.

The recently proposed SPD-ANC [17] is a hybrid ANC method where controller weights are updated using a least mean square (LMS) algorithm, and the secondary path and its reverse are modeled using two pre-trained time-domain CRNs. It is essentially an LMS based adaptive ANC method with the nonlinear distortions in the secondary path modeled by deep learning.

The CRN based frequency-domain method employs a CRN for complex spectral mapping and works by estimating the real and imaginary spectrograms of a canceling signal from the real and imaginary spectrograms of the reference signal [14]. The CRN has an encoder-decoder architecture with a two-layer grouped LSTM between them. The model sizes (the number of trainable parameters within a model) of the CRN and ARN based deep ANC methods are around 8.8 million and 15.9 million, respectively. Their multiply-accumulate (MAC) operations are 1.82 G and 5.24 G, respectively, for processing a 3-second noise.

C. Performance Metric

Normalized mean square error (NMSE) is used to evaluate the noise attenuation performance of the proposed method. NMSE is a widely used metric for ANC evaluations and it is defined as

$$\text{NMSE} = 10\log_{10}\left[\sum_{n=1}^{M} e^2(n) / \sum_{n=1}^{M} d^2(n)\right] \qquad (6)$$

where M is the total number of samples in a time-domain signal. NMSE values are typically below zero, and a lower value indicates better noise attenuation. The results shown in this paper are the average NMSE of 100 test signals.

V. Evaluation Results and Comparisons

A. Deep ANC With Shorter Frame Sizes

We first evaluate the performance of the proposed deep ANC model using different frame sizes. All ANC methods are tested with untrained noises in both linear ($\eta^2 = \infty$) and nonlinear ($\eta^2 = 0.5$ and $\eta^2 = 0.1$) situations and the comparison results are provided in Table I. Frame length and frame shift are connected by a dash, and the latter is set to half of the former. The corresponding algorithmic latency is shown inside the parentheses.

It can be seen that the performance of FxLMS degrades in the presence of nonlinear distortions. THF-FxLMS and the hybrid SPD-ANC yield better noise attenuation for nonlinear ANC. As expected, the optimal FxLMS solution with ground-truth secondary path and nonlinear distortion obtains the best noise attenuation among the adaptive ANC methods. Deep learning based methods are effective for noise attenuation in both linear and nonlinear situations and generalize well to untrained noises. Using shorter frame sizes results in lower algorithmic latencies. As mentioned previously, a longer frame size is usually used in frequency-domain methods to ensure acceptable frequency resolution. For CRN based deep ANC, which is a frequency-domain method, shortening frame size leads to worse ANC performance, e.g., noise attenuation level drops by more than 1 dB when the frame size is reduced from 20 ms to 4 ms. ARN based time-domain ANC consistently outperforms all the comparison methods, and it is even more advantageous for low-latency ANC as reducing frame size does not affect the performance.

We provide spectrograms of the outputs obtained using ARN based method with 4 ms frame size under different untrained noises in Fig. 9 to illustrate its noise attenuation performance. The first row of each panel shows the spectrogram of a primary noise, and the second row shows the residual noise (error signal) obtained after ANC. It is observed that the proposed method is capable of achieving wideband noise attenuation and its performance generalizes well to untrained noises.

B. Deep ANC With Delay-Compensated Training

This subsection investigates the performance of deep ANC with delay-compensated training. We start by comparing the performance of ARN and CRN based methods for noise attenuation with predicted noise. We use 20 ms frame size and 10 ms frame shift and train the deep ANC models to predict canceling signal for 5 ms, 10 ms, 15 ms, and 20 ms ahead. The results are given in Table II, and the corresponding algorithmic latencies are provided inside parentheses. Not surprisingly, the noise attenuation performance drops with the increase of prediction length. The table shows that ARN based time-domain deep ANC is more efficient at predicting noise; for example its noise attenuation performance drops only by 0.95 dB when predicting babble noise 15 ms ahead while the corresponding performance drop is 2.61 dB for the CRN based method. In the most challenging case of no algorithmic latency, the ARN based model exhibits a significant performance drop compared

TABLE I
AVERAGE NMSE (DB) OF TRADITIONAL ALGORITHMS AND DEEP ANC MODELS WITH DIFFERENT FRAME SIZES AND FRAME SHIFTS. ALGORITHMIC LATENCY OF EACH MODEL IS PROVIDED INSIDE THE PARENTHESES

Nonlinearity		Linear ($\eta^2 = \infty$)				Nonlinear ($\eta^2 = 0.5$)				Nonlinear ($\eta^2 = 0.1$)			
Noise		Babble	Engine	SSN	Factory	Babble	Engine	SSN	Factory	Babble	Engine	SSN	Factory
FxLMS		-6.04	-6.78	-5.95	-5.88	-4.32	-5.26	-4.38	-4.73	-3.37	-4.54	-3.46	-1.67
THF-FxLMS		–	–	–	–	-6.02	-6.70	-5.98	-5.86	-5.97	-6.55	-5.94	-5.75
SPD-ANC		-6.96	-8.75	-6.75	-6.70	-6.81	-8.55	-6.50	-6.51	-6.72	-8.49	-6.33	-6.40
Optimal FxLMS		-7.29	-9.29	-7.04	-7.16	-7.22	-9.21	-6.96	-7.14	-7.12	-9.00	-6.85	-7.10
CRN	20 ms – 10 ms (20 ms)	-10.58	-12.87	-11.36	-10.66	-10.54	-12.79	-11.30	-10.57	-10.37	-12.42	-11.03	-10.14
	16 ms – 8 ms (16 ms)	-10.43	-11.91	-10.76	-10.02	-10.39	-11.84	-10.70	-9.95	-10.24	-11.44	-10.40	-9.65
	4 ms – 2 ms (4 ms)	-9.51	-10.28	-10.25	-9.07	-9.48	-10.24	-10.21	-9.00	-9.41	-10.01	-9.95	-8.67
ARN	20 ms – 10 ms (20 ms)	-11.32	-12.67	-11.74	-11.24	-11.29	-12.65	-11.70	-11.18	-11.26	-12.62	-11.51	-10.74
	16 ms – 8 ms (16 ms)	-11.57	-12.87	-12.20	-11.72	-11.55	-12.87	-12.16	-11.66	-11.54	-12.90	-11.95	-11.03
	4 ms – 2 ms (4 ms)	-11.57	-11.96	-11.68	-11.49	-11.56	-11.95	-11.68	-11.46	-11.57	-12.01	-11.62	-11.08

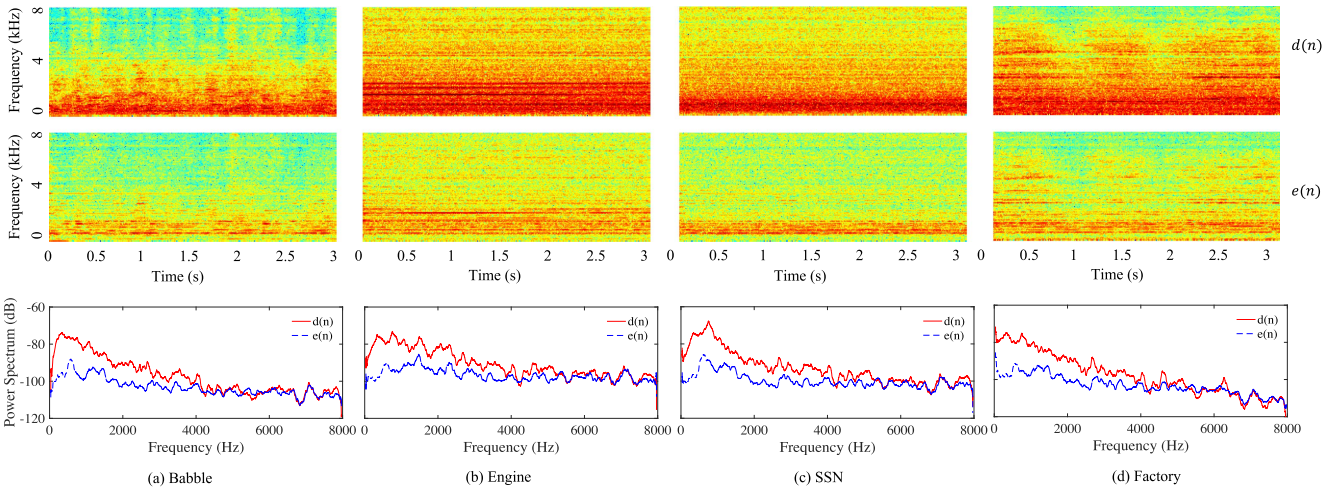

Fig. 9. Spectrograms and power spectra of test results for four noises. The first and second rows of each panel show the spectrograms of primary noise and output of deep ANC, respectively, and third row their power spectra. ARN-based ANC is implemented with frame size and frame shift set to 4 ms and 2 ms, respectively.

TABLE II
AVERAGE NMSE (DB) OF DEEP ANC MODELS WITH DELAY-COMPENSATED TRAINING TO PREDICT DIFFERENT NOISE LENGTHS. THE VALUE INSIDE THE PARENTHESES PROVIDES ALGORITHMIC LATENCY (IN MS) OF THE CORRESPONDING MODEL

Noise		Babble	Engine	SSN	Factory
CRN	No prediction (20 ms)	-10.58	-12.87	-11.36	-10.66
	Predicting 5 ms (15 ms)	-9.40	-10.89	-9.85	-9.35
	Predicting 10 ms (10 ms)	-8.76	-9.69	-9.10	-8.57
	Predicting 15 ms (5 ms)	-7.97	-9.41	-8.34	-7.39
	Predicting 20 ms (0 ms)	-7.18	-7.90	-7.81	-7.06
ARN	No prediction (20 ms)	-11.32	-12.67	-11.74	-11.24
	Predicting 5 ms (15 ms)	-10.80	-11.91	-11.51	-10.81
	Predicting 10 ms (10 ms)	-10.57	-12.14	-11.56	-10.81
	Predicting 15 ms (5 ms)	-10.37	-11.35	-10.68	-10.58
	Predicting 20 ms (0 ms)	-7.62	-8.44	-8.51	-7.88

to the case of 5 ms latency, although it still yields higher noise attenuation than the CRN based model.

There is a tradeoff between prediction length and ANC performance. To examine the ability of the proposed method for noise prediction, we gradually increase the prediction length and train multiple ARN models for active noise control. The frame size and frame shift of ARN are set to 4 ms and 2 ms, respectively. We vary the value of K and train the ARN based model to cancel primary noise with different time advances. The prediction results are shown in Fig. 10 with the algorithmic latency of each model provided inside the parentheses. It can be observed that the noise attenuation performance drops gradually with the increase of prediction length. Predicting 6 ms in advance, which reduces the algorithmic latency to -2 ms, still achieves good NMSE values. Predicting more than 6 ms ahead results in considerable performance drop compared to no prediction. We can conclude that, the proposed delay-compensated training strategy effectively reduces the algorithmic latency of deep ANC with acceptable levels of ANC performance degradation.

C. Deep ANC With Revised Overlap-Add

Deep ANC with revised OLA is evaluated in this part. First, we investigate the effects of revised OLA on ANC performances and provide the results using the original OLA method and two variations of it in Table III. The frame size and frame shift are set to 20 ms and 5 ms, respectively. We revise the OLA method by setting a portion (5 ms) of or the entire (15 ms) overlapping part to zero, with the corresponding algorithmic latencies of 15 ms and 5 ms, respectively. For both CRN and ARN based methods, using revised OLA leads to lower algorithmic latencies with small performance drops, and the ARN based method consistently outperforms the CRN based method. Having more

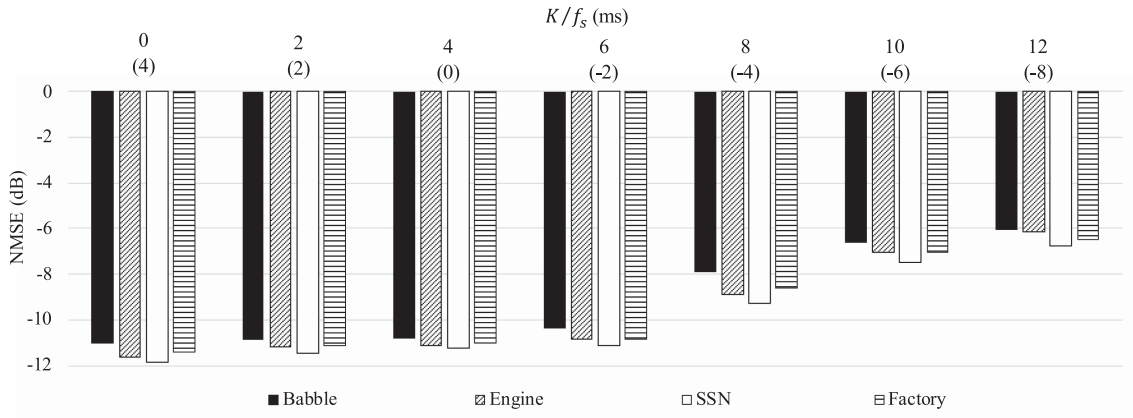

Fig. 10. Average NMSE (dB) for ARN based deep ANC with delay-compensated training to predict different noise lengths. The value inside the parentheses provides algorithmic latency (in ms) of the corresponding model.

TABLE III
AVERAGE NMSE (DB) OF DEEP ANC MODELS USING REVISED OLA WITH DIFFERENT ALGORITHMIC LATENCIES. THE VALUE INSIDE THE PARENTHESES PROVIDES ALGORITHMIC LATENCY (IN MS) OF THE CORRESPONDING MODEL

Noise		Babble	Engine	SSN	Factory
CRN	Original OLA (20 ms)	-11.00	-12.10	-11.46	-10.92
	Setting first 5 ms to 0 (15 ms)	-10.38	-11.78	-10.69	-10.18
	Setting first 15 ms to 0 (5 ms)	-10.07	10.36	-10.41	-9.76
ARN	Original OLA (20 ms)	-11.88	-13.20	-12.41	-12.27
	Setting first 5 ms to 0 (15 ms)	-11.20	-11.98	-11.89	-11.52
	Setting first 15 ms to 0 (5 ms)	-11.03	-12.09	-11.73	-11.46

TABLE IV
AVERAGE NMSE (DB) OF DEEP ANC MODELS USING REVISED OLA WITH NO OVERLAP. THE NMSE RESULTS OF MODELS WITH THE SAME ALGORITHMIC LATENCY BUT USING PREDICTING ARE PROVIDED

Noise		Babble	Engine	SSN	Factory
CRN 20 ms – 10 ms	No prediction (20 ms)	-10.58	-12.87	-11.36	-10.66
	Predicting 10 ms (10 ms)	-8.76	-9.69	-9.10	-8.57
	Revised OLA (10 ms)	-9.85	-11.13	-10.37	-9.86
ARN 20 ms - 10 ms	No prediction (20 ms)	-11.32	-12.67	-11.74	-11.24
	Predicting 10 ms (10 ms)	-10.57	-12.14	-11.56	-10.81
	Revised OLA (10 ms)	-10.91	-12.15	-11.62	-10.98
ARN 4 ms – 2 ms	No prediction (4 ms)	-11.57	-11.96	-11.68	-11.49
	Predicting 2 ms (2 ms)	-10.89	-11.18	-11.46	-11.12
	Revised OLA (2 ms)	-10.91	-11.30	-11.62	-11.22

overlapping samples is beneficial for signal resynthesis and results in better noise attenuation. It is observed that the revised OLA method with fewer overlapping samples has slightly worse performance but lower latency. Compared to the original OLA method, setting the first 15 ms of each frame to zeros results in a 1.22 dB less noise attenuation for CRN based deep ANC, and 0.86 dB less attenuation for ARN based deep ANC.

We will use the revised OLA with no overlapping samples as the default setting in the following experiments.

Second, we compare the effectiveness of delay-compensated training and revised OLA for reducing algorithmic latency. The results of deep ANC models with the same algorithmic latency but using different strategies are provided in Table IV.

Using revised OLA achieves a little better performance than noise prediction for reducing algorithmic latency by the same amount, especially for the CRN based method. Where using revised OLA produces more than 1 dB noise attenuation than using delay-compensated training with 10 ms algorithmic latency. However, this does not indicate that revised OLA is superior to delay-compensated training since the former can at most reduce the algorithmic latency to the length of frame shift J while there is no restriction for delay-compensated training. However, revised OLA and delay-compensated training can be combined to further reduce algorithmic latency.

TABLE V
AVERAGE NMSE (DB) OF DEEP ANC MODELS TRAINED USING DIFFERENT STRATEGIES TO ACHIEVE ZERO OR NEGATIVE ALGORITHMIC LATENCIES

Noise		Babble	Engine	SSN	Factory
CRN	20 ms – 10 ms predicting 20 ms (0 ms)	-7.18	-7.90	-7.81	-7.06
	20 ms – 10 ms revised OLA + predicting 10 ms (0 ms)	-7.68	-8.27	-8.33	-7.64
	20 ms – 5 ms revised OLA + predicting 5 ms (0 ms)	-9.06	-9.93	-9.57	-9.13
	4 ms – 2 ms predicting 4 ms (0 ms)	-9.09	-8.94	-9.38	-8.46
	4 ms – 2 ms revised OLA + predicting 2 ms (0 ms)	-9.70	-9.48	-9.98	-9.83
ARN	20 ms – 10 ms predicting 20 ms (0 ms)	-7.62	-8.44	-8.51	-7.88
	20 ms – 10 ms revised OLA + predicting 10 ms (0 ms)	-7.75	-8.46	-8.52	-7.98
	20 ms – 5 ms revised OLA + predicting 5 ms (0 ms)	-10.04	-10.30	-10.85	-10.28
	4 ms – 2 ms predicting 4 ms (0 ms)	-10.80	-11.12	-11.23	-11.00
	4 ms – 2 ms revised OLA + predicting 2 ms (0 ms)	-10.85	-11.23	-11.58	-11.16
	4 ms – 2 ms revised OLA + predicting 4 ms (-2 ms)	-10.62	-11.05	-11.28	-10.93

D. Low-Latency Deep ANC

We now evaluate deep ANC using different frame sizes, frame shifts, and combine different training strategies to achieve low-latency ANC. Table V shows the results with zero and even negative algorithmic latency. The first five rows give the results of the CRN based model and the last 6 rows provide the results of the ARN based model. We know from Table I that without considering algorithmic latency, using longer frame sizes results in better ANC performance for the CRN based method. However, for pure prediction cases (the first and the

TABLE VI
AVERAGE NMSE (DB) OF ARN BASED LOW-LATENCY DEEP ANC TESTED ON RECORDED NOISES

	ARN (0 ms)	ARN (-2 ms)
NRIVER	-9.44	-9.31
OMEETING	-10.20	-10.15
DLIVING	-10.52	-10.20
PRESTO	-10.04	-10.16
SPSQUARE	-10.78	-10.41
TMETRO	-10.27	-9.69

fourth rows of Table V), using smaller frame sizes is preferred to achieve zero algorithmic latency since the total samples that need to be predicted are substantially fewer. A similar trend is observed in the results of the ARN based model. From the first two rows of CRN and ARN, we observe that combining revised OLA and delay-compensated training is more efficient for reaching zero algorithmic latency than relying on noise prediction only. Given the same frame length, we find that a shorter frame shift is desirable for achieving zero latency. This is because the algorithmic latency is reduced to the length of frame shift with the help of revised OLA, and using smaller frame shifts requires predicting fewer samples, a relatively easier task than using larger shifts. Moreover, smaller frame shifts result in more overlaps between input frames, which is helpful for estimation. Using the ARN based model with 4 ms frame size, 2 ms frame shift, revised OLA, and delay-compensated training (predicting 2 ms) achieves the best ANC performance among all these models with 0 ms algorithmic latency. The average NMSE is -11.20 dB in this case, and there is only 0.47 dB performance drop compared to the average NMSE of the ARN based model with the same frame size and frame shift but without using revised OLA and delay-compensated training (the last row of Table I). A clear way to measure progress is by comparing to the model in our previous study [14] (i.e. row 1 of Table II), and the algorithmic latency is reduced from 20 ms to 0 ms with no performance degradation.

Finally, non-stationary noises from the DEMAND corpus [55] are used to test the performance of deep ANC in realistic conditions. The DEMAND dataset has six categories of noises, and we choose one noise from each category to represent distinct environments:

- NRIVER noise: from the "Nature" category, recorded besides a creek of running water.
- OMEETING noise, from the "Office" category, recorded in a meeting room.
- DLIVING noise: from the "Domestic" category, recorded inside a living room.
- PRESTO noise: from the "Public" category, recorded in a university restaurant at lunchtime.
- SPSQUARE noise: from the "Street" category, recorded in a public town square with many tourists.
- TMETRO noise: from the "Transportation" category, recorded in a subway.

Table VI gives the average NMSE results, which show that the proposed deep ANC works well for recorded noises in different realistic environments.

In general, ARN based time-domain ANC is effective for low-latency deep ANC. Combining ARN with the proposed strategies leads to zero or even negative algorithmic latency without significantly affecting ANC performance. Zero or negative algorithmic latency would be impossible for traditional ANC methods, and goes a long way to alleviating the causality constraint, facilitating the design of ANC systems, and expanding the scope of ANC applications.

VI. CONCLUSION

This study focuses on low-latency deep ANC. We have proposed a time-domain deep ANC method based on attentive recurrent networks with smaller frame sizes to reduce algorithmic latency. Augmented with delay-compensated training and revised OLA, the algorithmic latency of deep ANC is substantially reduced, which largely alleviates the causality constraint of ANC systems and facilitates ANC design. The performance of low-latency deep ANC using different strategies has been evaluated, and the combination of these strategies leads to zero and even negative algorithmic latency. Future research will investigate practical issues of device implementation. For example, DNN model compression has been shown to be effective for reducing model sizes dramatically without significant performance degradation [56]. We also plan to evaluate the proposed system using measured acoustic paths. In addition, we will extend the proposed low-latency strategies to other audio processing tasks such as speech enhancement and speaker separation.

REFERENCES

[1] G. C. Goodwin, E. I. Silva, and D. E. Quevedo, "Analysis and design of networked control systems using the additive noise model methodology," *Asian J. Control*, vol. 12, pp. 443–459, 2010.
[2] J. Cheer and S. J. Elliott, "Multichannel control systems for the attenuation of interior road noise in vehicles," *Mech. Syst. Signal Process.*, vol. 60, pp. 753–769, 2015.
[3] S. M. Kuo, S. Mitra, and W. S. Gan, "Active noise control system for headphone applications," *IEEE Trans. Control Syst. Technol.*, vol. 14, no. 2, pp. 331–335, Mar. 2006.
[4] J. F. Wilby, "Aircraft interior noise," *J. Sound Vib.*, vol. 190, pp. 545–564, 1996.
[5] T. Murao, C. Shi, W. S. Gan, and M. Nishimura, "Mixed-error approach for multi-channel active noise control of open windows," *Appl. Acoust.*, vol. 127, pp. 305–315, 2017.
[6] L. Liu, S. Gujjula, and S. M. Kuo, "Multi-channel real time active noise control system for infant incubators," in *Proc. IEEE Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2009, pp. 935–938.
[7] S. M. Kuo and D. R. Morgan, "Active noise control: A tutorial review," *Proc. IEEE*, vol. 87, no. 6, pp. 943–973, Jun. 1999.
[8] S. Elliott, I. Stothers, and P. Nelson, "A multiple error LMS algorithm and its application to the active control of sound and vibration," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 10, pp. 1423–1434, Oct. 1987.
[9] F. Agerkvist, "Modelling loudspeaker non-linearities," in *Proc. Audio Eng. Soc. Conf.: 32nd Int. Conf. DSP Loudspeakers*, 2007.
[10] W. Klippel, "Tutorial: Loudspeaker nonlinearities—causes, parameters, symptoms," *J. Audio Eng. Soc.*, vol. 54, pp. 907–939, 2006.
[11] M. H. Costa, J. C. M. Bermudez, and N. J. Bershad, "Stochastic analysis of the filtered-X LMS algorithm in systems with nonlinear secondary paths," *IEEE Trans. Signal Process.*, vol. 50, no. 6, pp. 1327–1342, Jun. 2002.
[12] B. Lam, W. S. Gan, D. Shi, M. Nishimura, and S. Elliott, "Ten questions concerning active noise control in the built environment," *Building Environ.*, vol. 200, 2021, Art. no. 107928.
[13] H. Zhang and D. L. Wang, "A deep learning approach to active noise control," in *Proc. Interspeech*, 2020, pp. 1141–1145.
[14] H. Zhang and D. L. Wang, "Deep ANC: A deep learning approach to active noise control," *Neural Netw.*, vol. 141, pp. 1–10, 2021.

[15] H. Zhang and D. L. Wang, "A deep learning method to multi-channel active noise control," in *Proc. Interspeech*, 2021, pp. 681–685.
[16] D. Shi, B. Lam, K. Ooi, X. Shen, and W. S. Gan, "Selective fixed-filter active noise control based on convolutional neural network," *Signal Process.*, vol. 190, 2022, Art. no. 108317.
[17] 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 Process. Lett.*, vol. 29, pp. 234–238, 2021.
[18] H. Zhang and D. L. Wang, "Deep MCANC: A deep learning approach to multi-channel active noise control," *Neural Netw.*, vol. 158, pp. 318–327, 2023.
[19] X. Kong and S. M. Kuo, "Study of causality constraint on feedforward active noise control systems," *IEEE Trans. Circuits Syst. II. Analog Digit. Signal Process.*, vol. 46, no. 2, pp. 183–186, Feb. 1999.
[20] L. Zhang and X. Qiu, "Causality study on a feedforward active noise control headset with different noise coming directions in free field," *Appl. Acoust.*, vol. 80, pp. 36–44, 2014.
[21] R. A. Burdisso, J. S. Vipperman, and C. R. Fuller, "Causality analysis of feedforward-controlled systems with broadband inputs," *J. Acoust. Soc. Amer.*, vol. 94, pp. 234–242, 1993.
[22] S. Kurczyk and M. Pawełczyk, "Fuzzy control for semi-active vehicle suspension," *J. Low Freq. Noise Vib. Act. Control.*, vol. 32, no. 3, pp. 217–225, 2013.
[23] S. Kurczyk and M. Pawełczyk, "Active noise control using a fuzzy inference system without secondary path modelling," *Arch. Acoust.*, vol. 39, 2014, pp. 243–248.
[24] S. Kurczyk and M. Pawełczyk, "Nonlinear structural acoustic control with shunt circuit governed by a soft-computing algorithm," *Arch. Acoust.*, vol. 43, no. 3, pp. 397–402, 2018.
[25] W. K. Tseng, B. Rafaely, and S. Elliott, "Combined feedback–feedforward active control of sound in a room," *J. Acoust. Soc. Amer.*, vol. 104, pp. 3417–3425, 1998.
[26] K. Iwai, S. Hase, and Y. Kajikawa, "Multichannel feedforward active noise control system with optimal reference microphone selectorbased on time difference of arrival," *Appl. Sci.*, vol. 8, 2018, Art. no. 2291.
[27] M. Parviainen, P. Pertilä, T. Virtanen, and P. Grosche, "Time-frequency masking strategies for single -channel low-latency speech enhancement using neural networks," in *Proc. IEEE 16th Int. Workshop Acoust. Signal Enhancement*, 2018, pp. 51–55.
[28] G. Naithani, G. Parascandolo, T. Barker, N. H. Pontoppidan, and T. Virtanen, "Low-latency sound source separation using deep neural networks," in *Proc. IEEE Glob. Conf. Signal Inf. Process.*, 2016, pp. 272–276.
[29] E. Friot, "Time-domain versus frequency-domain effort weighting in active noise control design," *J. Acoust. Soc. Amer.*, vol. 141, pp. EL11–EL15, 2017.
[30] 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," *Appl. Sci.*, vol. 8, 2018, Art. no. 2313.
[31] D. Shi, W.-S. Gan, B. Lam, R. Hasegawa, and Y. Kajikawa, "Feedforward multichannel virtual-sensing active control of noise through an aperture: Analysis on causality and sensor-actuator constraints," *J. Acoust. Soc. Amer.*, vol. 147, pp. 32–48, 2020.
[32] S. W. Fu, Y. Tsao, X. Lu, and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," in *Proc. IEEE Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2017, pp. 006–012.
[33] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2018, pp. 5069–5073.
[34] S. W. Fu, T. W. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 9, pp. 1570–1584, Sep. 2018.
[35] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.
[36] A. Pandey and D. Wang, "A new framework for CNN-based speech enhancement in the time domain," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 7, pp. 1179–1188, Jul. 2019.
[37] A. Pandey and D. Wang, "Self-attending RNN for speech enhancement to improve cross-corpus generalization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, no. 3, pp. 1374–1385, Mar. 2022.
[38] H. Zhang, A. Pandeyand, and D. L. Wang, "Attentive recurrent network for low-latency active noise control," in *Proc. Interspeech*, 2022, pp. 956–960.
[39] R. Crochiere, "A weighted overlap-add method of short-time Fourier analysis/synthesis," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 1, pp. 99–102, Feb. 1980.
[40] S. U. N. Wood and J. Rouat, "Unsupervised low latency speech enhancement with RT-GCC-NMF," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 2, pp. 332–346, May 2019.
[41] S. Merity, "Single headed attention RNN: Stop thinking with your head," 2019, *arXiv:1911.11423*.
[42] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[43] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.
[44] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.
[45] J. Chen, Y. Wang, S. E. Yoho, D. L. Wang, and E. W. Healy, "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," *J. Acoust. Soc. Amer.*, vol. 139, pp. 2604–2612, 2016.
[46] 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 Commun.*, vol. 12, pp. 247–251, 1993.
[47] J. Cheer, "Active control of the acoustic environment in an automobile cabin," Ph.D. dissertation, Univ. Southampton, Southampton, U.K., 2012.
[48] P. N. Samarasinghe, W. Zhang, and T. D. Abhayapala, "Recent advances in active noise control inside automobile cabins: Toward quieter cars," *IEEE Signal Process. Mag.*, vol. 33, no. 6, pp. 61–73, Nov. 2016.
[49] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, pp. 943–950, 1979.
[50] S. Ghasemi, R. Kamil, and M. H. Marhaban, "Nonlinear THF-FxLMS algorithm for active noise control with loudspeaker nonlinearity," *Asian J. Control*, vol. 18, pp. 502–513, 2016.
[51] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of adam and beyond," 2019, *arXiv:1904.09237*.
[52] F. Yang, J. Guo, and J. Yang, "Stochastic analysis of the filtered-x LMS algorithm for active noise control," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2252–2266, 2020.
[53] W. Chen and Z. Zhang, "Nonlinear adaptive learning control for unknown time-varying parameters and unknown time-varying delays," *Asian J. Control*, vol. 13, pp. 903–913, 2011.
[54] D. Huang and J. X. Xu, "Discrete-time adaptive control for nonlinear systems with periodic parameters: A lifting approach," *Asian J. Control*, vol. 14, pp. 373–383, 2012.
[55] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings," in *Proc. Meetings Acoust.*, 2013, Art. no. 035081.
[56] K. Tan and D. L. Wang, "Towards model compression for deep learning based speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1785–1794, 2021.