Hao Zhang[a,*], DeLiang Wang[a,b]

[a]Department of Computer Science and Engineering, Ohio State University, Columbus, OH 43210-1277 USA
[b]Center for Cognitive and Brain Sciences, Ohio State University, Columbus, OH 43210-1277 USA

Abstract

Traditional active noise control (ANC) methods are based on adaptive signal processing with the least mean square algorithm as the foundation. They are linear systems and do not perform satisfactorily in the presence of nonlinear distortions. In this paper, we formulate ANC as a supervised learning problem and propose a deep learning approach, called deep ANC, to address the nonlinear ANC problem. The main idea is to employ deep learning to encode the optimal control parameters corresponding to different noises and environments. A convolutional recurrent network (CRN) is trained to estimate the real and imaginary spectrograms of the canceling signal from the reference signal so that the corresponding anti-noise can eliminate or attenuate the primary noise in the ANC system. Large-scale multi-condition training is employed to achieve good generalization and robustness against a variety of noises. The deep ANC method can be trained to achieve active noise cancellation no matter whether the reference signal is noise or noisy speech. In addition, a delay-compensated strategy is introduced to solve the potential latency problem of ANC systems. Experimental results show that deep ANC is effective for wideband noise reduction and generalizes well to untrained noises. Moreover, the proposed method can achieve ANC within a quiet zone and is robust against variations in reference signals.

Keywords: Active noise control, Deep learning, Deep ANC, Loudspeaker nonlinearity, Quiet zone

1. Introduction

Active noise control is a noise cancellation methodology based on the principle of superposition of acoustic signals, i.e. two superposed waveform signals cancel each other when they have the same amplitude but the opposite phase. The goal of ANC systems is to generate an anti-noise with the same amplitude and opposite phase of the primary (unwanted) noise to cancel the primary noise (Goodwin et al., 2010). ANC differs from passive noise control, e.g. by using sound-absorbing barriers like an earplug, and noise removal in signal enhancement where noise is removed by processing a noisy signal like noisy speech (Wang & Chen, 2018). Fundamentally, ANC requires to predict both the amplitude and phase of a noise signal at a given point in space ahead of time. While signal amplitude may be steady over time, signal phase changes all the time at any spatial location due to the nature of acoustic waves (Hartmann, 2004). Thus ANC is a very challenging problem, and in practice it can attenuate only low-frequency stationary noises.

ANC has attracted increasing attention in research and industrial applications over the past few decades. There are two kinds of ANC systems: feedforward and feedback (Kuo & Morgan, 1999). A typical feedforward ANC system is shown in Fig. 1, and it consists of a reference microphone, a canceling loudspeaker, and an error microphone. The active noise controller takes the reference signal and error signal, sensed by the reference microphone and error microphone, respectively, as inputs

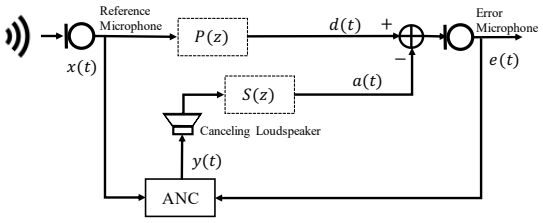


Figure 1: Diagram of a single-channel feedforward ANC system, where $P(z)$ and $S(z)$ denote the frequency responses of the primary path and secondary path, respectively.

to adapt the controller so that the canceling signal generated can superpose with the primary noise at the location to be silenced. Feedback ANC uses only an error sensor to adapt the controller and is simpler to implement. However, it is not as effective as feedforward ANC when dealing with broad-band noise cancellation because feedback ANC does not use the information from the reference signal.

Traditionally, an active noise controller is implemented using adaptive filters that optimize filter characteristics by minimizing an error signal (Manolakis et al., 2000). Filtered-x least mean square (FxLMS) and its extensions are the most widely used active noise controllers due to their simplicity, robustness and relatively low computational load. The FxLMS algorithm alleviates the effect of the secondary path by filtering the reference signal, $x(t)$, with an estimate of the secondary path before feeding it to the controller (Elliott et al., 1987); see Fig. 1. The secondary path is usually estimated separately as a finite impulse response filter (FIR) beforehand. However, nonlinear distor-

*Corresponding author
Email addresses: zhang.6720@osu.edu (Hao Zhang), dwang@cse.ohio-state.edu (DeLiang Wang)

as amplifiers and loudspeakers. The linear adaptive approach fails to identify the secondary path accurately in the presence of nonlinearities. Consequently, the inaccurately estimated secondary path deteriorates the overall noise cancellation performance. Costa et al. (2002) present a statistical analysis of the FxLMS behavior when the secondary path includes a nonlinear element and conclude that a small nonlinearity can have a significant impact on the adaptive filter behavior.

Many adaptive nonlinear ANC algorithms have been proposed in the literature to address nonlinear distortions (Lashkari, 2006; Tan & Jiang, 2001; Ghasemi et al., 2016; Kuo & Wu, 2005; Das & Panda, 2004; Tobias & Seara, 2005; Napoli & Piroddi, 2009). The Volterra expansion has been shown to be effective for modeling soft or weak nonlinearities (Lashkari, 2006) and a truncated second-order Volterra based FxLMS algorithm has been proposed for feedforward active noise control in the presence of nonlinear distortions (Tan & Jiang, 2001; Guo et al., 2018). Napoli & Piroddi (2009) utilize the polynomial Nonlinear AutoRegressive model with eXogenous variables (NARX) to identify controller structure for more efficient and reliable nonlinear ANC. Nonlinear FxLMS and the tangential hyperbolic function based FxLMS (THF-FxLMS) are introduced to handle the nonlinearities of the ANC system by modeling the secondary path as a saturation-type nonlinearity (Ghasemi et al., 2016). Other algorithms such as bilinear FxLMS (Kuo & Wu, 2005), filtered-s LMS (Das & Panda, 2004), leaky FxLMS (Tobias & Seara, 2005) have also been investigated to address nonlinearity. However, their performance is limited in the presence of strong nonlinearities.

Neural networks have been introduced to address nonlinear ANC (George & Panda, 2013), considering their ability in handling nonlinear relations. A multilayer perceptron (MLP) network is introduced in Snyder & Tanaka (1995) for active control of vibrations, wherein the weights of the neural network are updated by using adaptive filtered-x backpropagation. Based on the nonlinear active control structure given in Snyder & Tanaka (1995), improved training algorithms are developed to increase the convergence speed and decrease the computational load of the training (Bouchard et al., 1999; Chang & Luoh, 2007). Panda & Das (2003) and Krukowicz (2010) use an efficient ANC structure based on functional link neural network to resolve the nonlinear effect in ANC. Other nonlinear adaptive models such as radial basis function networks (Tokhi & Wood, 1997), fuzzy neural networks (Zhang et al., 2006), and recurrent neural networks (Bambang, 2008) have been developed to further improve the ANC performance. These neural network architectures for nonlinear ANC utilize online adaptation or training to obtain an optimal controller and thus should be regarded as adaptive algorithms.

Deep learning is capable of modeling complex nonlinear relationships and can potentially play an important role in addressing nonlinear ANC problems. To be suitable for real-world applications, an ANC system must be able to attenuate a variety of noises and cope with the variations in acoustic environments. Traditional ANC systems are adaptive and condition training would be required to expose ANC to a large variety of noises and variations during training. A deep learning model trained this way could potentially generalize to untrained noises and environments.

In this paper, we propose a new approach to address ANC, particularly nonlinear ANC problems. Our approach, named deep ANC, employs a deep learning model trained to encode the optimal control parameters corresponding to different noise sources. Considering that ANC is inherently sensitive to both the magnitude and phase of the anti-noise, we use complex spectral mapping to estimate both magnitude and phase responses of the ANC output simultaneously (Williamson et al., 2016; Fu et al., 2017; Tan & Wang, 2019b). During training, a CRN (Tan & Wang, 2019a) is trained to estimate the real and imaginary spectrograms of a canceling signal from the reference signal. The subsequent anti-noise is obtained by passing the canceling signal through a loudspeaker and secondary path (see Fig. 1). Finally, the error signal is used to calculate the loss function for training the CRN model. To the best of our knowledge, this study is the first attempt to formulate ANC as a supervised learning problem and use deep learning to address it.

From the methodological perspective, deep ANC can be more advantageous than traditional ANC algorithms. Besides attenuating noise signals, deep ANC can be trained to attenuate the noise components of a noisy speech signal and let the underlying speech pass through. Namely, deep ANC in principle is able to maintain the target signal embedded in noise by selectively canceling the noise components of the noisy signal; the target signal does not have to be speech, and it can be other kinds, such as music. This advantage could dramatically expand the scope of ANC applicability. In addition, we introduce a delay-compensated training strategy to tackle a shortcoming of frequency-domain ANC algorithms: processing latency.

Besides ANC at a given spatial location, a more useful but more challenging task is to perform ANC within a small spatial zone, i.e., to generate a quiet zone. The deep ANC method can be trained in an RIR (room impulse response) independent way to generate such a quiet zone.

A preliminary version of this study is recently published (Zhang & Wang, 2020). Compared to the conference version, this paper provides a wider range of evaluations with more noises, signal-to-noise ratios (SNRs), and untrained speakers. In addition, the robustness of deep ANC against variations that occur in reference signals is investigated and new comparisons are made with other nonlinear ANC methods.

The remainder of this paper is organized as follows. Section II introduces the signal model of active noise control. Section III presents the deep ANC approach. Evaluation metrics and experimental setup are given in Section IV. Section V shows the evaluation results and comparisons. Section VI concludes the paper.

2.1. Signal Model

As is shown in Fig. 1, the primary path and secondary path correspond to the acoustic responses from the reference microphone and the canceling loudspeaker, respectively, to the error microphone, and their frequency responses are denoted as $P(z)$ and $S(z)$, respectively. The reference signal $x(t)$ is picked up by a reference microphone and passed through the active noise controller to get the canceling signal $y(t)$. The canceling signal is then passed through the canceling loudspeaker and the secondary path to produce the anti-noise $a(t)$. The corresponding error signal sensed by the error microphone is defined as

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

where t is the time index, $d(t)$ is the primary signal received by the error microphone, $w(t)$ represents the active noise controller, $f_{\text{LS}}\{\cdot\}$ denotes the function of the loudspeaker, $*$ denotes linear convolution, and the superscript T indicates transpose. Furthermore, $p(t)$ and $s(t)$ denote the impulse responses of the primary and secondary path, respectively.

Active noise control aims to generate an anti-noise so as to cancel the primary noise. Traditionally, this is accomplished by using adaptive algorithms to estimate the digital filter $W(z)$ so that the mean squared error is minimized. The FxLMS algorithm and its variations work by estimating a secondary path first and then placing the estimated filter $\hat{S}(z)$ in the reference signal path to compensate for the effect of the secondary path. The secondary path is usually estimated offline during an initial stage in ANC applications and the performance of traditional ANC methods depends largely on the accuracy of $\hat{S}(z)$.

2.2. Deep Learning for Active Noise Control

Ignoring the function of the loudspeaker, the z-transform of (1) can be written as

$$E(z) = [P(z) - S(z)W(z)]X(z) \tag{2}$$

Assuming that the residual error is completely attenuated after the convergence of an adaptive filter, the optimal solution of the adaptive filter can be represented by the following transfer function

$$W^o(z) = \frac{P(z)}{S(z)} \tag{3}$$

This means that the ANC system has to model $P(z)$ and the inverse of $S(z)$, simultaneously, to achieve the optimal performance (Kuo & Morgan, 1999). However, an inverse of $S(z)$ does not necessarily exist and the direct estimation of the adaptive filer $W(z)$ can be complicated for traditional adaptive algorithms, letting alone the nonlinear distortions introduced by the loudspeaker.

Different from traditional ANC methods that need to estimate the secondary path and adaptive filter individually, deep ANC uses supervised learning and trains a deep neural network to directly approximate the optimal controller $W^o(z)$ in order to minimize the error signal under different situations. A diagram of deep ANC is shown in Fig. 2. The overall goal is to estimate a canceling signal from the reference signal so that the corresponding anti-noise cancels the primary noise. In the proposed method, we use the reference signal as input and set the ideal anti-noise as the training target. To achieve complete noise cancellation, the ideal anti-noise should be the same as the primary noise. During training, the output of deep ANC is treated as an "intermediate product" and the estimate of the anti-noise is generated by passing deep ANC output through the loudspeaker and the secondary path. The loss function is calculated from the error signal.

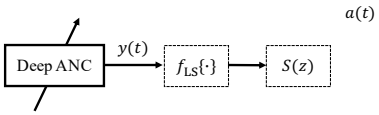


Figure 2: Diagram of the deep ANC approach, where $f_{\text{LS}}\{\cdot\}$ denotes the function of the canceling loudspeaker and $y(t)$ the canceling signal (output of deep ANC).

Formulating ANC as a supervised learning problem is nontrivial. There are two conceptual obstacles to such a formulation. First, it is not straightforward to define what the training target should be for a deep neural network (DNN). Although the ideal canceling signal for attenuating a primary noise is known, it cannot be used directly as the desired output of the DNN due to the existence of the loudspeaker and the secondary path (see Fig. 2). Second, the primary and secondary paths can be time-varying and the transfer function that the DNN needs to approximate can be different for different acoustic conditions. This seems to imply that a supervised learning model needs to predict a one-to-many mapping, an impossible job. These obstacles may explain why ANC has not been approached from the deep learning standpoint. However, as detailed in the next section, we have access to the ideal anti-noise to supervise DNN training, and the DNN can be trained to estimate, for a given input, some average of the different outputs for different scenarios. With these observations, ANC can be formulated as a deep learning task.

3. Deep ANC Method

3.1. Feature Extraction and Training Target

The reference signal $x(t)$ is sampled at 16 kHz and divided into 20-ms frames with a 10-ms overlap between consecutive frames. Then a 320-point short time Fourier transform (STFT) is applied to each time frame to produce the real and imaginary spectrograms of $x(t)$, which are denoted as $X_r(m, c)$ and $X_i(m, c)$, respectively, within a T-F unit at time m and frequency c. The proposed CRN for deep ANC is shown in Fig. 3 and it takes $X_r(m, c)$ and $X_i(m, c)$ as input features for complex spectral mapping.

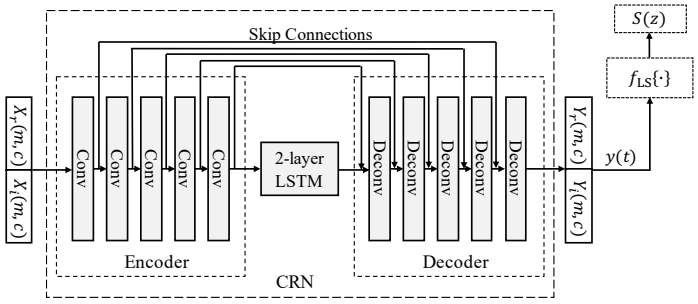


Figure 3: Diagram of CRN based deep ANC. Conv blocks denote convolutional layers and Deconv blocks denote deconvolutional layers. Skip connections connect layers at the same level. The inputs and outputs of CRN are defined in the complex STFT domain.

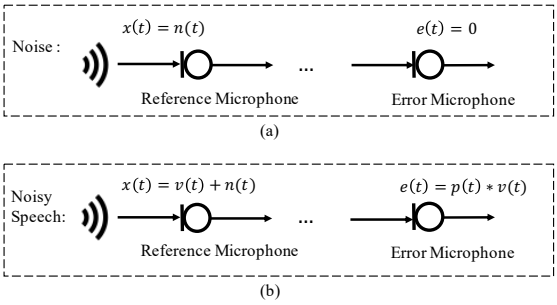


Figure 4: Illustration of the training strategies for deep ANC when reference signal is (a) noise, and (b) noisy speech.

To attenuate the primary noise at the location of the error microphone, deep ANC uses the ideal anti-noise (the primary noise) as the training target. The CRN is trained to map from the real and imaginary spectrograms of the reference signal to those of the canceling signal, $Y_r(m, c)$ and $Y_i(m, c)$. This is different from the methods that estimate only the magnitude spectrogram and use the phase spectrogram of the input signal to generate the estimated waveform output. We choose complex spectral mapping because of the importance of phase in active noise control. The complex spectrogram of the canceling signal goes through the inverse Fourier transform to derive a waveform signal $y(t)$. The anti-noise, which can be regarded as an estimate of the training target, is then generated by passing the canceling signal through the loudspeaker and secondary path.

3.2. Two Training Strategies and Loss Functions

In real-world applications, ANC applications may need to handle cases when the reference signal is noisy speech. Taking noise-canceling headphones as an example, the reference microphone on the headphones may pick up voice when someone is talking in the vicinity of the user. The reference signal in this case is a mixture of speech and primary noise. In this case, ANC should ideally allow the speech signal to pass through while suppressing the primary noise.

Deep ANC can be trained to achieve noise cancellation no matter whether the reference signal is noise or noisy speech, by using proper training data and loss functions. Fig. 4 shows two training strategies for the deep ANC method:

- **Deep ANC trained with noise**: The model trained this way aims to cancel any noise received at the reference microphone. To achieve this, we use noise signal $n(t)$ as the reference signal and train deep ANC to completely eliminate the primary noise. The loss function is defined as:

$$L_n = \frac{\sum_{t=1}^{L} e^2(t)}{L} \tag{4}$$

  where L is the length of the noise signal, and $e(t)$ is defined in (1).

- **Deep ANC trained with noisy speech**: The deep ANC model is trained to cancel surrounding noise while preserving speech signal. The reference signal used to train this deep ANC system is a mixture of noise $n(t)$ and speech $v(t)$, and the corresponding primary signal $d(t)$ is

$$\begin{aligned} d(t) &= p(t) * [v(t) + n(t)] \\ &= p(t) * v(t) + p(t) * n(t) \end{aligned} \tag{5}$$

  where $p(t) * n(t)$ and $p(t) * v(t)$ are, respectively, the noise and speech components of the primary signal. In order to attenuate only noise components and let speech pass through, the training target is set to the noise component, $p(t) * n(t)$, and the ideal error signal then is equivalent to $p(t) * v(t)$. The loss function used for training this deep ANC system is defined as:

$$L_{ns} = \frac{\sum_{t=1}^{L} [e(t) - p(t) * v(t)]^2}{L} \tag{6}$$

3.3. Learning Machine

Deep ANC employs a CRN for complex spectral mapping (Tan & Wang, 2019a). Besides its previous use for complex spectral mapping and strong speech enhancement performance, the CRN exhibits higher parameter efficiency and is suitable for real-time processing. The CRN has an encoder-decoder architecture, as shown in Fig. 3, where the encoder and decoder

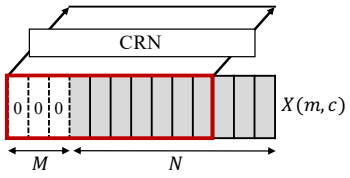


Figure 5: Diagram of the delay-compensated training strategy. Gray blocks denote the original signals and blocks inside the thick boxes denote the expanded signals for the delay-compensated strategy. The CRN block denotes the convolutional recurrent network used for deep ANC.

comprise five convolutional layers and five deconvolutional layers, respectively. Between them are two recurrent LSTM (long short-term memory) layers with a group strategy (Gao et al., 2018), where the group number is set to 2. The encoder-decoder structure is designed in a symmetric way where the number of kernels progressively increases in the encoder and decreases in the decoder. To aggregate the spectral context, a stride of two is adopted along the frequency dimension in all convolutional and deconvolutional layers. Therefore, the frequency dimensionality of feature maps is halved layer by layer in the encoder and doubled layer by layer in the decoder to ensure that the output has the same shape as the input. Skip connections are utilized in the CRN so that the output of each encoder layer is concatenated to the input of the corresponding decoder layer. In the CRN, all convolutions and deconvolutions are causal, so that the system does not use future information and is thus suited for real-time implementation. A detailed description of the CRN architecture is provided in Tan & Wang (2019a).

We employ exponential linear units (ELUs) (Clevert et al., 2015) in all convolutional and deconvolutional layers except for the output layer, where linear activation is used for spectrogram estimation. Moreover, we utilize batch normalization (Ioffe & Szegedy, 2015) right after each convolution or deconvolution and before activation. The model is trained using the AMSGrad optimizer (Reddi et al., 2019) with a learning rate of 0.001 for 30 epochs.

3.4. Delay-Compensated Training

The proposed approach uses real and imaginary spectrograms as the input and output, and it can thus be regarded as a frequency-domain ANC algorithm. However, frequency-domain ANC algorithms incur a time delay equal to the frame length of STFT (Yang et al., 2018). This delay may violate the causality constraint of ANC, considered a shortcoming for frequency-domain ANC algorithms. Many approaches have been proposed to reduce this delay, which is not easy to be completely eliminated (Kim et al., 1994; Kuo et al., 2008; Park et al., 2001; Bendel et al., 2001; Rout et al., 2015).

We propose a delay-compensated training strategy for deep ANC in order to address this delay problem. The main idea is to train the model to predict the canceling signal a few frames in advance. A diagram of this strategy is shown in Fig. 5, where N denotes the total number of frames in an input signal, M denoting. Then the first N frames of the expanded input are used as the new input signal to train the model. The new input is a concatenation of M frames of zeros and the first $N - M$ frames of the original input, while the target signal is kept unchanged, hence equivalent to using the input signal to predict M frames of the target in advance. With 20-ms frames and 10-ms frame shift, delay-compensated training saves $10 \times M$ ms for active noise control. Therefore, this strategy can in principle solve the latency problem of frequency-domain ANC systems.

4. Experimental Setup

4.1. Performance Metrics

Performance of the proposed method is evaluated in terms of normalized mean squared error (NMSE), short-time objective intelligibility (STOI) (Taal et al., 2011) and perceptual evaluation of speech quality (PESQ) (Rix et al., 2001).

The power of the error signal in an ANC system is usually used as a quality metric of noise attenuation. In this paper, we use NMSE in dB to measure the performance of ANC systems:

$$\text{NMSE} = 10 \log_{10} \frac{\sum_{t=1}^{L} e^2(t)}{\sum_{t=1}^{L} d^2(t)} \tag{7}$$

where L is the length of signal. The value of NMSE is usually below zero and a lower value indicates better noise attenuation.

STOI and PESQ are used to measure the intelligibility and quality of denoised speech received at the error microphone, respectively, when the reference signal is noisy speech. They are obtained by comparing the error signal $e(t)$ with the speech component of the primary signal, $p(t) * v(t)$. The range of the STOI score is typically from 0 to 1. The range of the PESQ score is from -0.5 to 4.5. A higher score is better.

4.2. Experimental Settings

To train a noise-independent model, we expose the ANC model to a large variety of noisy environments in the training stage (Chen et al., 2016) and use 10000 non-speech environmental sounds (noises) from a sound-effect library (http://www.sound-ideas.com) to create the training set. Engine noise, factory noise, babble noise and speech-shaped noise (denoted as "SSN") from NOISEX-92 (Varga & Steeneken, 1993) are used for testing. Note that the testing noises are unseen during the training stage, and hence evaluate the generalization ability of the proposed method.

The physical structure of an ANC system is usually modeled as a rectangular enclosure (Kestell, 2000; Tarabini & Roure, 2008) and many studies have shown the effectiveness of ANC systems for noise canceling in enclosed rooms (Sommerfeldt et al., 1995; Cheer, 2012; Samarasinghe et al., 2016). In this study, we simulate a rectangular enclosure of size 3 m $\times$ 4 m $\times$ 2 m (width $\times$ length $\times$ height) and use the image method (Allen & Berkley, 1979) to generate room impulse responses for the primary and secondary paths of an ANC system. The

Noise type	Engine			Factory			Babble			SSN		
η^2	∞	0.5	0.1	∞	0.5	0.1	∞	0.5	0.1	∞	0.5	0.1
FxLMS	-6.78	-5.26	-4.54	-5.88	-4.73	-1.67	-6.04	-4.32	-3.37	-5.95	-4.38	-3.46
THF-FxLMS	–	-6.70	-6.55	–	-5.86	-5.75	–	-6.02	-5.97	–	-5.98	-5.94
CRN-n	**-11.07**	**-10.98**	**-10.60**	**-9.58**	**-9.50**	**-9.17**	**-9.49**	**-9.45**	**-9.27**	**-9.90**	**-9.83**	**-9.56**
CRN-n(-1)	-9.60	-9.53	-9.25	-8.47	-8.42	-8.19	-8.80	-8.76	8.62	-8.97	-8.92	-8.69
CRN-n(-2)	-7.93	-7.89	-7.72	-6.97	-6.94	-6.81	-7.00	-7.00	-6.89	-7.50	-7.47	7.32

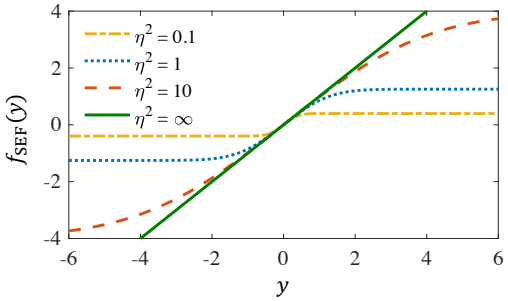


Figure 6: Scaled error function for various values of η^2.

reference microphone is located at the position (1.5, 1, 1) m, the canceling loudspeaker is located at (1.5, 2.5, 1) m and the error microphone at (1.5, 3, 1) m. This is a typical scenario where the primary noise source is located far from the walls for easy access (Lau & Tang, 2000). For training, five different reverberation times (T60s): 0.15 s, 0.175 s, 0.2 s, 0.225 s, and 0.25 s, are used for generating RIRs and the length of the RIRs is set to 512. Two RIRs, one for primary path and the other one for secondary path, are generated with each T60. For testing, we use the RIRs with reverberation time 0.2 s as the default test RIRs, and the RIRs generated with untrained T60s are also used to test the generalization ability of deep ANC.

Saturation effects produced by a loudspeaker are the most significant source of nonlinearity present in an ANC system (Costa et al., 2002; Ghasemi et al., 2016). In ANC studies of loudspeaker saturation (Agerkvist, 2007; Klippel, 2006; Tobias & Seara, 2006; Bershad, 1990), this nonlinearity is usually represented by the scaled error function (SEF) (Tobias & Seara, 2006):

$$f_{\mathrm{SEF}}(y) = \int_0^y e^{-\frac{z^2}{2\eta^2}} dz \tag{8}$$

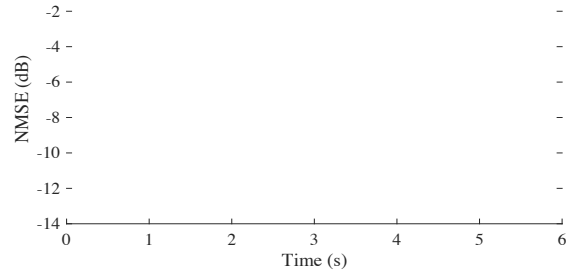
where y is the input to the loudspeaker (see Fig. 1), and η^2 represents the strength of nonlinearity. It models a commonly found saturation type nonlinearity, e.g. sound level saturation limited by loudspeaker size. The SEF becomes linear as η^2 tends to infinity and becomes a hard limiter as η^2 tends to zero. To investigate the robustness of the proposed method against nonlinear distortions, four loudspeaker functions are used during the training stage: $\eta^2 = 0.1$ (severe nonlinearity), $\eta^2 = 1$ (moderate nonlinearity), $\eta^2 = 10$ (soft nonlinearity), and $\eta^2 = \infty$ (linear). Fig. 6 plots the SEF with these η^2 values. For testing, we use both trained and untrained loudspeaker functions.

The deep ANC is trained to handle cases when the reference signal is either noise or noisy speech. To achieve this, we generate 20000 training signals and 100 test signals for each case. Each noise signal is created by randomly cutting a 6-second signal from the 10000 noise signals. The speech signal used to generate noisy speech is obtained from the TIMIT dataset (Lamel et al., 1989) by randomly choosing 200 speakers (100 male speakers and 100 female speakers). Each chosen speaker has 10 utterances in the TIMIT corpus, and 7 of them are used for training and the remaining 3 for testing. To create a noisy speech signal, utterances from a randomly selected speaker are mixed with random noise cuts from the 10000 noises at a SNR randomly chosen from [5, 10, 15, 20] dB. The primary signal $d(t)$ is generated by convolving the a reference signal with a randomly selected RIR for the primary path. The anti-noise $a(t)$ is generated by passing the corresponding canceling signal $y(t)$ successively through a randomly chosen loudspeaker function and a randomly selected RIR for the secondary path. In the following experiments, we use CRN-n and CRN-ns to denote the deep ANC model trained with noise and noisy speech, respectively.

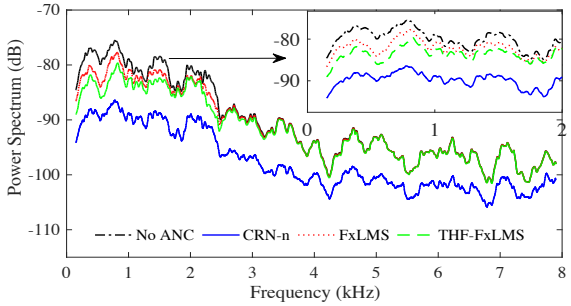
4.3. Comparison Methods

The deep ANC method is compared with FxLMS and THF-FxLMS in linear and nonlinear situations. FxLMS works by modeling the secondary path in terms of an FIR filter and utilizing the estimated model to adapt the filter for the ANC controller. FxLMS achieves good noise attenuation when the secondary path is a linear system. However, it fails to identify the secondary path accurately when there is nonlinear distortion in the system. Nonlinear models for the secondary path are utilized in active noise control to account for nonlinear distortions. THF-FxLMS is a recently proposed algorithm for nonlinear ANC (Ghasemi et al., 2016). It incorporates the tangent hyperbolic function (THF) to model the saturation effect of the loudspeaker and then apply the estimated degree of nonlinearity in the design of the ANC controller. As shown in Ghasemi et al. (2016), THF-FxLMS outperforms FxLMS and the second-order Volterra algorithm for noise attenuation in situations with nonlinear distortions.

Both FxLMS and THF-FxLMS are adaptive algorithms and can be used to cancel different types of noise. However, their performance is sensitive to control parameters such as the step size and filter length. Appropriate step sizes are needed to achieve good performance when exposed to different noises and environments. The step sizes of FxLMS and THF-FxLMS in our experiments are chosen heuristically for different noises ac-



(a) Normalized mean squared error



(b) Power spectrum

Figure 7: Noise attenuation achieved for engine noise with loudspeaker nonlinearity $\eta^2 = 0.1$: (a) normalized mean squared error, (b) power spectrum.

cording to the criteria given in Chen & Zhang (2011) and Huang & Xu (2012) to ensure stable updating and good noise attenuation. The filter length of the comparison methods is set to 512, which is equal to the length of the primary and secondary paths.

In addition, we consider another nonlinear ANC setup in Section 5.5 where deep ANC is compared with a Volterra filter based method and a MLP based method.

5. Evaluation Results and Comparisons

5.1. Performance of Deep ANC Trained with Noise

We first evaluate the performance of the deep ANC model trained with noise. The proposed methods and traditional ANC algorithms are tested with four types of untrained noises in a linear system ($\eta^2 = \infty$) and two nonlinear systems ($\eta^2 = 0.5$, $\eta^2 = 0.1$). Table 1 shows the average NMSE of 100 testing signals, where CRN-n(-1), and CRN-n(-2) denote the model trained with the delay-compensated strategy to predict 1 and 2 frames in advance, respectively. The step size used to update FxLMS with respect to engine noise, factory noise, babble noise and SSN is set to 0.05, 0.4, 0.3, 0.4, respectively, and the step size to update THF-FxLMS is set to 0.05, 0.4, 0.3, 0.4, respectively, for the four noises. It is apparent from this table that FxLMS is capable of attenuating different noises but its performance degrades when it comes to nonlinear ANC. THF-FxLMS models the secondary path as a nonlinear system and it achieves good noise attenuation in different nonlinear cases. The deep ANC models outperform the comparison algorithms

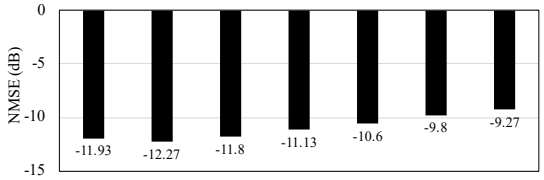


Figure 8: Average NMSE for CRN-n with engine noise, $\eta^2 = 0.1$, and untrained RIRs with different T60s. The NMSE result for the trained T60 of 0.2 s is included as a reference.

in both linear and nonlinear cases and generalize well to untrained noises and untrained nonlinearity ($\eta^2 = 0.5$). As expected, CRN-n performs the best among the deep ANC models. Using the delay-compensated strategy to predict one or even two frames still obtains good levels of noise attenuation, higher than FxLMS and THF-FxLMS, while the overall performance drops gradually as prediction length increases.

We plot the NMSE and the power spectrum curves in Fig. 7 for further comparison. Power spectrum measures signal power with respect to frequency and is used here to show relative noise attenuation achieved at different frequencies. The results in this figure are obtained in the situations with engine noise and nonlinearity $\eta^2 = 0.1$. It can be seen that the deep ANC method consistently outperforms the comparison methods. As illustrated in Fig. 7(b) , the proposed method achieves wideband noise reduction, while the comparison methods are only effective for noise attenuation at low frequencies. It is well known that traditional ANC is restricted to low frequencies (Kuo & Morgan, 1999; Samarasinghe et al., 2016), due to factors such as convergence and latency. As a result, narrow band noise or low-pass filtered noise are usually used as input. Deep ANC is effective for both low- and high-frequency noises. Note that the test noises used in this study are wideband, which is part of the reason why the amount of noise removal for the comparison methods is lower in Table 1 than typically reported in the literature.

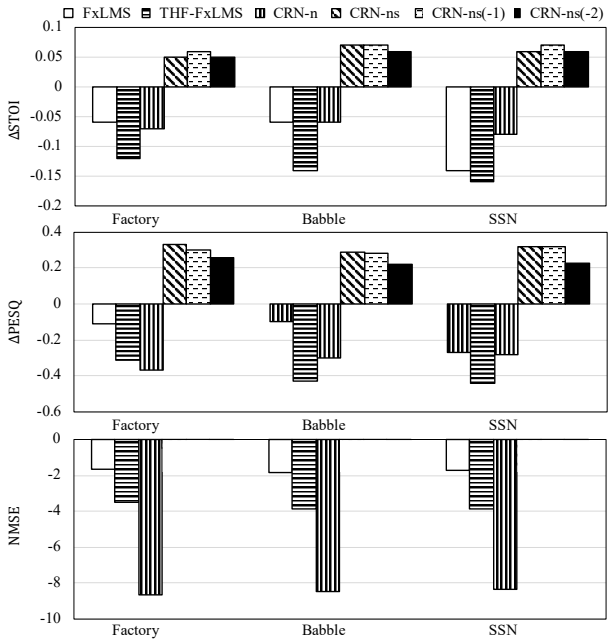
Figure 8 gives the average NMSE of deep ANC when tested with RIRs generated with different T60 values. It shows that the performance of deep ANC generalizes well to untrained RIRs.

5.2. Performance of Deep ANC Trained with Noisy Speech

This subsection evaluates the performance of deep ANC when the reference signal is noisy speech. Comparison results of different methods in situations with noise and noisy speech at different SNR levels are given in Table 2, where CRN-ns(-1) and CRN-ns(-2) denote the models trained with the delay-compensated strategy to predict 1 and 2 frames in advance, respectively. The step size to update FxLMS with engine noise at the SNR of 5,15, and 20 dB is set to 0.01, 0.05, 0.01, respectively, and the step size to update THF-FxLMS is set to 0.01, 0.01, 0.01, respectively, at the three SNR levels. The results given in this table are obtained with engine noise and a nonlinear system with $\eta^2 = 0.1$. "Unprocessed" denotes the results when there is no ANC, and the STOI and PESQ values of un-

and engine noise at different SNR levels.

	Noise only	SNR = 5 dB			SNR = 15 dB			SNR = 20 dB		
Models	NMSE	STOI	PESQ	NMSE	STOI	PESQ	NMSE	STOI	PESQ	NMSE
Unprocessed	0	0.79	1.95	0	0.94	2.61	0	0.97	2.96	0
FxLMS	-4.54	0.71	1.84	-2.30	0.71	1.90	-0.36	0.72	1.98	-0.07
THF-FxLMS	-6.55	0.69	1.73	-3.89	0.74	1.92	-1.58	0.78	2.12	-0.99
CRN-n	**-10.60**	0.72	1.71	**-8.75**	0.83	2.02	**-8.66**	0.85	2.10	**-8.67**
CRN-ns	-10.00	0.84	**2.26**	–	**0.96**	**3.00**	–	**0.98**	**3.32**	–
CRN-ns(-1)	-8.31	**0.85**	2.24	–	**0.96**	2.95	–	**0.98**	3.28	–
CRN-ns(-2)	-7.04	0.84	2.14	–	**0.96**	2.87	–	**0.98**	3.21	–



Unprocessed	Factory	Babble	SSN
STOI	0.83	0.75	0.78
PESQ	2.23	2.01	1.95

Figure 9: Performance comparison in noisy speech situations with different noises at SNR = 5 dB and loudspeaker nonlinearity $\eta^2 = 0.1$. The unprocessed average STOI and PESQ values are given in the table below the plots.

processed signals are obtained by comparing the primary signal $d(t)$ with the underlying speech component in it. The second column in Table 2 shows the NMSE values when tested with noise signals. It can be seen that the performance of CRN-ns is comparable to that of CRN-n when tested in the noise only situation even though the former model is trained with noisy speech. For situations with noisy speech, the overall performance is dropped in terms of NMSE since speech is the main component in noisy speech with positive SNRs. The CRN-n model still has the best performance and the NMSE values at all SNR levels are below −8.6 dB. That is, the CRN-n model treats noisy speech as "general noise" and it is capable of attenuating noise, as well as noisy speech. CRN-ns trained with noisy speech aims to remove the noise component of noisy speech and the error signal corresponds to an estimate of clean speech.

We use STOI and PESQ to evaluate the performance of preserving the speech component. As shown in the table, CRN-ns improves objective intelligibility and quality of speech, due to its ability to selectively attenuate noise. For example, there is around 0.05 STOI and 0.3 PESQ improvement at the SNR level of 5 dB. Performances of CRN-ns(-1) and CRN-ns(-2) are comparable to that of CRN-ns, with a small decrease in terms of PESQ. Traditional methods and CRN-n focus on minimizing error signal (attenuating reference signal), and therefore distort the speech component as reflected by substantially lower STOI and PESQ values than unprocessed noisy speech.

The deep ANC method is further tested with factory noise, babble noise and SSN at 5 dB SNR to show its robustness to different noises. To clearly show the improvement in terms of STOI and PESQ, we define ΔSTOI and ΔPESQ as the difference in these metrics introduced by ANC. Fig. 9 plots these values and NMSE values. It can be seen that ΔSTOI and ΔPESQ values for FxLMS, THF-FxLMS and CRN-n are all below zero. The models trained with noisy speech (CRN-ns, CRN-ns(-1) and CRN-ns(-2)) generalize well to untrained noises and are capable of selectively attenuating the noise components of noisy speech, hence improving objective speech intelligibility and quality.

Waveforms and spectrograms of CRN-ns with nonlinearity $\eta^2 = 0.1$ are shown in Fig. 10, where the first row shows the results when tested with engine noise and the second row when tested with noisy speech with engine noise at 5 dB SNR. It is evident that the deep ANC system trained with noisy speech (CRN-ns) can not only attenuate the noise component in the noisy speech, but also cancel the noise when there is no speech in the reference signal.

5.3. Quiet Zone

So far we have focused on noise attenuation at a given point in space. A more challenging and desirable task would be to achieve active noise control within a small spatial zone, called quiet zone (Zhu et al., 2020; Kuo et al., 2004). To achieve a quiet zone, deep ANC can be trained in an RIR-independent way by exposing the model to a variety of RIRs sampled within a small zone during training. To be specific, we simulate the quiet zone as a sphere with a radius of 5 cm which is appropriate, for example, for around one ear of a driver inside a vehicle. We randomly select 100 points within the sphere as the locations of the error microphone and generate 100 pairs of RIRs for primary and secondary paths by using the image method

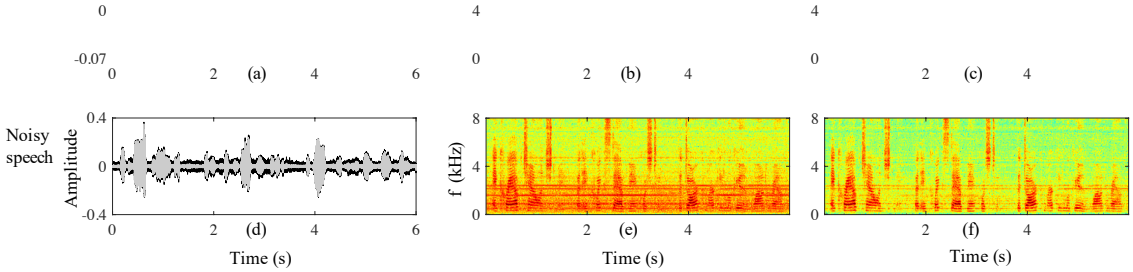


Figure 10: Waveforms and spectrograms of the CRN-ns model at noise only and noisy speech cases with engine noise, 5 dB SNR, and loudspeaker nonlinearity $\eta^2 = 0.1$: (a) and (d) are waveforms of primary signal (black lines) and error signal (gray lines), (b) and (e) are spectrograms of the primary signals, and (c) and (f) are spectrograms of the error signals.

Table 3: Average NMSE (dB), STOI and PESQ for deep ANC trained to generate a quiet zone with 5 cm radius for different noises and different SNR levels. The nonlinear distortion is $\eta^2 = 0.1$.

Model	CRN-n				CRN-ns (Engine)					
Noise	Engine	Factory	Babble	SSN	SNR = 5 dB		SNR = 15 dB		SNR = 20 dB	
	NMSE	NMSE	NMSE	NMSE	STOI	PESQ	STOI	PESQ	STOI	PESQ
Unprocessed	0	0	0	0	0.79	1.95	0.94	2.63	0.97	2.96
$r = 0$	-9.44	-9.87	-10.33	-10.09	0.85	2.16	0.96	2.88	0.98	3.21
$r = 2$	-9.49	-10.18	-10.55	-10.39	0.85	2.17	0.96	2.90	0.98	3.20
$0 \leq r \leq 5$	-8.32	-8.97	-9.42	-9.13	0.85	2.16	0.96	2.88	0.98	3.19

Table 4: Robustness of deep ANC to variations in reference signal with nonlinear distortion $\eta^2 = 0.1$.

$\eta^2 = 0.1$		Noise Only	Noisy Speech		
Cases:		NMSE	STOI	PESQ	NMSE
SNR Change	Unprocessed	0	0.91	2.64	0
	CRN-n	-10.60	0.76	2.02	-8.44
	CRN-ns	-10.00	0.93	2.97	–
Noise Change	Unprocessed	0	0.81	2.06	0
	CRN-n	-10.05	0.73	1.77	-8.69
	CRN-ns	-9.93	0.86	2.39	–
Noise Mixture	Unprocessed	0	0.80	2.01	0
	CRN-n	-9.46	0.73	1.78	-8.84
	CRN-ns	-9.52	0.86	2.37	–
Untrained Speakers	Unprocessed	–	0.78	1.96	0
	CRN-n	–	0.71	1.72	-8.37
	CRN-ns	–	0.83	2.22	–

(Allen & Berkley, 1979). Twenty thousand training signals for noise only and noisy speech cases are created with these 100 pairs of RIRs in each case, and the CRN-n and CRN-ns models are trained with these data. Three test sets, with 100 signals in each, are generated to evaluate the performance of these models.

The results are given in Table 3, where r denotes the distance from the zone center. For the case of "$r = 0$", test signals are generated by placing the error microphone at the center of the sphere. For "$r = 2$", the test set is generated by placing the microphone within the sphere, 2 cm away from the center point. For the case of "$0 \leq r \leq 5$", we randomly place the error microphone at 10 different points within the sphere and use the corresponding 10 pairs of RIRs to create the test set. The CRN-n model produces 8.32 dB NMSE for noisy speech with engine noise within the sphere. The CRN-ns model obtains 0.06 and 0.21 improvement, respectively, in terms of STOI and PESQ for the engine noise at 5 dB SNR within the sphere. Similar amounts of NMSE, STOI, and PESQ improvements are observed for other test conditions in Table 3. Generally speaking, the deep ANC models trained in this way achieve substantial noise attenuation at any point within this sphere while preserving speech, hence generating a quiet zone.

5.4. Robustness of deep ANC

In ANC applications, many variations occur in reference signals such as SNR, noise type, and multiple noises existing simultaneously in the reference signal. To test the robustness of the proposed methods against these variations, we evaluate the model trained in Sections 5.1 and 5.2 in four cases. First, the SNR level of reference signal is changed from 5 dB to 20 dB (with engine noise) after 3 seconds, the middle point of a reference signal. Second, the noise type in reference signal is changed from engine noise to factory noise after 3 seconds. Third, the reference signal is a mixture of engine noise and factory noise. The SNR levels for the second and third case are set to 5 dB. Fourth, we randomly select 20 untrained speakers (10 male and 10 female) from the TIMIT dataset and create 100 test signals (with engine noise and 5 dB SNR) to evaluate the performance of deep ANC in an untrained speaker condition. The results are given in Table 4. "Noisy Speech" results are obtained by using noisy speech as the reference signal, and "Noise Only" results by using the noise component of noisy speech as the reference signal. Results shown in this table demonstrate the strong robustness of the deep ANC approach.

setup.

NMSE	Engine	Factory	Babble	SSN
VFxLMS	-15.48	-13.75	-15.86	-13.58
FxMLP	-17.01	-14.96	-16.13	-15.57
CRN-n	-19.63	-20.32	-20.15	-20.01

5.5. Comparison using different nonlinear ANC setup

We consider a different nonlinear ANC setup in this subsection and compare the performance of deep ANC with a Volterra filter based method and a neural network based method. The experiments are carried out using the setup in Guo et al. (2018) and Zhou & DeBrunner (2007).

The primary path is modeled by a Volterra series and the relationship between the primary noise $d(t)$ and the reference signal $x(t)$ is defined as (Zhou & DeBrunner, 2007; Guo et al., 2018)

$$\begin{aligned} d(t) =& x(t) + 0.8x(t-1) + 0.3x(t-2) + 0.4x(t-3) \\ & - 0.8x(t)x(t-1) + 0.9x(t)x(t-2) + 0.7x(t)x(t-3) \end{aligned} \tag{9}$$

The secondary path is modeled as the nonlinear-linear (NL) structure introduced in Zhou & DeBrunner (2007). In the NL model, the anti-noise $a(t)$ is obtained by passing the canceling signal $y(t)$ successively through a nonlinear model, denoted as N, and an FIR filter that is denoted as $l(z)$ in the z-domain,

$$N[y(t)] = 3.3 \tanh[0.3y(t)] \tag{10}$$

$$l(z) = 1 + 0.2z^{-1} + 0.05z^{-2} \tag{11}$$

We have implemented the adaptive Volterra controller using the FxLMS structure (VFxLMS) introduced in Guo et al. (2018) and Tan & Jiang (2001). The active noise controller and the secondary path of the VFxLMS algorithm are modeled by using second-order Volterra filters with a memory size of 10 and the step sizes are set as given in Guo et al. (2018).

The neural network based method for nonlinear ANC is an extension of the FxLMS algorithm with the controller modeled as an MLP (Chang & Luoh, 2007; Zhou et al., 2005). It is denoted as FxMLP and the weights of the MLP are updated adaptively using gradient descent (Chang & Luoh, 2007). The MLP has 6 inputs, 2 hidden layers with 12 neurons in each layer, and 1 neuron in the output layer. The hidden layer activation functions are sigmoidal and the last layer is linear.

As for deep ANC, we generate another 20000 training signals from the 10000 noises and retrain the CRN-n model. These training signals are generated following the description in Section 4.2 except that the primary and the secondary paths are replaced with the ones presented in (9), (10), and (11).

CRN-n, VFxLMS, and FxMLP are evaluated with respect to the four types of noise used before with 100 noise signals generated for each condition. The comparison results are given in Table 5. It can be seen that all of these methods are capable of attenuating noise for the nonlinear ANC setup, and deep ANC consistently outperforms the other two methods.

In this paper, we have introduced the deep ANC approach to active noise control. A convolutional recurrent network is employed to estimate a canceling signal from the reference signal so as to remove or attenuate the primary noise. Using proper training data and loss functions, the deep ANC system can be trained to not only cancel noise, but also selectively cancel the noise component of noisy speech. We have also proposed a delay-compensated training strategy to tackle the latency problem of frequency-domain ANC methods. In addition, the proposed method is capable of achieving ANC within a spatial zone. Systematic evaluations with NMSE, STOI and PESQ show the effectiveness and robustness of the deep ANC model for noise attenuation in noise only and noisy speech situations and the model generalizes well to different acoustic variations.

The deep ANC approach has major advantages over traditional methods. It has the intrinsic ability of modeling nonlinearities unavoidable in ANC systems. Deep ANC is flexible in terms of training target, e.g., it can be trained to achieve noise cancellation in noisy speech and even noisy music. A quiet zone can be generated by using a single canceling loudspeaker, whereas adaptive filter methods need multiple loudspeakers. Unlike traditional methods, deep ANC is effective for wideband noise removal. Finally, in addition to address the latency difficulty in frequency-domain algorithms, the delay compensation strategy significantly expands the range of causality in ANC.

Future work includes exploring time-domain methods for deep ANC, assessing robustness of deep ANC to RIR changes caused by changing error microphone position, and extending deep ANC to a multi-channel version. Furthermore, practical issues such as computational complexity and device implementation will be considered in future research.

Acknowledgements

We thank Jingli Xie, Xiao-Lei Zhang, and Wen Zhang for discussions in the early stage of this work. This research was supported in part by two NIDCD grants (R01 DC012048 and R01 DC015521) and the Ohio Supercomputer Center.

References

Agerkvist, F. (2007). Modelling loudspeaker non-linearities. In *Audio Engineering Society Conference: 32nd International Conference: DSP For Loudspeakers*.

Allen, J. B., & Berkley, D. A. (1979). Image method for efficiently simulating small-room acoustics. *Journal of the Acoustical Society of America*, *65*, 943–950.

Bambang, R. T. (2008). Adjoint EKF learning in recurrent neural networks for nonlinear active noise control. *Applied Soft Computing*, *8*, 1498–1504.

Bendel, Y., Burshtein, D., Shalvi, O., & Weinstein, E. (2001). Delayless frequency domain acoustic echo cancellation. *IEEE Transactions on Speech and Audio Processing*, *9*, 589–597.

Bershad, N. J. (1990). On weight update saturation nonlinearities in LMS adaptation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *38*, 623–630.

Bouchard, M., Paillard, B., & Le Dinh, C. T. (1999). Improved training of neural networks for the nonlinear active control of sound and vibration. *IEEE Transactions on Neural Networks*, *10*, 391–401.

348–356.

Cheer, J. (2012). *Active control of the acoustic environment in an automobile cabin*. Ph.D. thesis University of Southampton.

Chen, J., Wang, Y., Yoho, S. E., Wang, D. L., & Healy, E. W. (2016). Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises. *Journal of the Acoustical Society of America*, *139*, 2604–2612.

Chen, W., & Zhang, Z. (2011). Nonlinear adaptive learning control for unknown time-varying parameters and unknown time-varying delays. *Asian Journal of Control*, *13*, 903–913.

Clevert, D.-A., Unterthiner, T., & Hochreiter, S. (2015). Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, .

Costa, M. H., Bermudez, J. C. M., & Bershad, N. J. (2002). Stochastic analysis of the filtered-x LMS algorithm in systems with nonlinear secondary paths. *IEEE Transactions on Signal Processing*, *50*, 1327–1342.

Das, D. P., & Panda, G. (2004). Active mitigation of nonlinear noise processes using a novel filtered-s LMS algorithm. *IEEE Transactions on Speech and Audio Processing*, *12*, 313–322.

Elliott, S., Stothers, I., & Nelson, P. (1987). A multiple error LMS algorithm and its application to the active control of sound and vibration. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *35*, 1423–1434.

Fu, S.-W., Hu, T.-y., Tsao, Y., & Lu, X. (2017). Complex spectrogram enhancement by convolutional neural network with multi-metrics learning. In *IEEE 27th International Workshop on Machine Learning for Signal Processing* (pp. 1–6).

Gao, F., Wu, L., Zhao, L., Qin, T., Cheng, X., & Liu, T.-Y. (2018). Efficient sequence learning with group recurrent networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 799–808).

George, N. V., & Panda, G. (2013). Advances in active noise control: A survey, with emphasis on recent nonlinear techniques. *Signal Processing*, *93*, 363–377.

Ghasemi, S., Kamil, R., & Marhaban, M. H. (2016). Nonlinear THF-FXLMS algorithm for active noise control with loudspeaker nonlinearity. *Asian Journal of Control*, *18*, 502–513.

Goodwin, G. C., Silva, E. I., & Quevedo, D. E. (2010). Analysis and design of networked control systems using the additive noise model methodology. *Asian Journal of Control*, *12*, 443–459.

Guo, X., Li, Y., Jiang, J., Dong, C., Du, S., & Tan, L. (2018). Sparse modeling of nonlinear secondary path for nonlinear active noise control. *IEEE Transactions on Instrumentation and Measurement*, *67*, 482–496.

Hartmann, W. M. (2004). *Signals, sound, and sensation*. Springer Science & Business Media.

Huang, D., & Xu, J.-X. (2012). Discrete-time adaptive control for nonlinear systems with periodic parameters: A lifting approach. *Asian Journal of Control*, *14*, 373–383.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, .

Kestell, C. D. (2000). *Active control of sound in a small single engine aircraft cabin with virtual error sensors*. Ph.D. thesis Adelaide University.

Kim, I.-S., Na, H.-S., Kim, K.-J., & Park, Y. (1994). Constraint filtered-x and filtered-u least-mean-square algorithms for the active control of noise in ducts. *Journal of the Acoustical Society of America*, *95*, 3379–3389.

Klippel, W. (2006). Tutorial: Loudspeaker nonlinearitiescauses, parameters, symptoms. *Journal of the Audio Engineering Society*, *54*, 907–939.

Krukowicz, T. (2010). Active noise control algorithm based on a neural network and nonlinear input-output system identification model. *Archives of Acoustics*, *35*, 191–202.

Kuo, S. M., & Morgan, D. R. (1999). Active noise control: a tutorial review. *Proceedings of the IEEE*, *87*, 943–973.

Kuo, S. M., & Wu, H.-T. (2005). Nonlinear adaptive bilinear filters for active noise control systems. *IEEE Transactions on Circuits and Systems I: Regular Papers*, *52*, 617–624.

Kuo, S. M., Wu, H.-T., Chen, F.-K., & Gunnala, M. R. (2004). Saturation effects in active noise control systems. *IEEE Transactions on Circuits and Systems I: Regular Papers*, *51*, 1163–1171.

Kuo, S. M., Yenduri, R. K., & Gupta, A. (2008). Frequency-domain delayless

Lamel, L. F., Kassel, R. H., & Seneff, S. (1989). Speech database development: Design and analysis of the acoustic-phonetic corpus. In *Speech Input/Output Assessment and Speech Databases*.

Lashkari, K. (2006). A novel Volterra-Wiener model for equalization of loudspeaker distortions. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings* (pp. V–V).

Lau, S., & Tang, S. (2000). Sound fields in a slightly damped rectangular enclosure under active control. *Journal of Sound and Vibration*, *238*, 637–660.

Manolakis, D. G., Ingle, V. K., Kogon, S. M. et al. (2000). *Statistical and adaptive signal processing: spectral estimation, signal modeling, adaptive filtering, and array processing*. McGraw-Hill Boston.

Napoli, R., & Piroddi, L. (2009). Nonlinear active noise control with NARX models. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*, 286–295.

Panda, G., & Das, D. P. (2003). Functional link artificial neural network for active control of nonlinear noise processes. In *2003 International Workshop on Acoustic Echo and Noise Control* (pp. 163–6).

Park, S. J., Yun, J. H., Park, Y. C., & Youn, D. H. (2001). A delayless subband active noise control system for wideband noise control. *IEEE Transactions on Speech and Audio Processing*, *9*, 892–899.

Reddi, S. J., Kale, S., & Kumar, S. (2019). On the convergence of Adam and beyond. *arXiv preprint arXiv:1904.09237*, .

Rix, A. W., Beerends, J. G., Hollier, M. P., & Hekstra, A. P. (2001). Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)* (pp. 749–752).

Rout, N. K., Das, D. P., & Panda, G. (2015). Computationally efficient algorithm for high sampling-frequency operation of active noise control. *Mechanical Systems and Signal Processing*, *56*, 302–319.

Samarasinghe, P. N., Zhang, W., & Abhayapala, T. D. (2016). Recent advances in active noise control inside automobile cabins: Toward quieter cars. *IEEE Signal Processing Magazine*, *33*, 61–73.

Snyder, S. D., & Tanaka, N. (1995). Active control of vibration using a neural network. *IEEE Transactions on Neural Networks*, *6*, 819–828.

Sommerfeldt, S. D., Parkins, J. W., & Park, Y. C. (1995). Global active noise control in rectangular enclosures. In *Proceedings of the INTER-NOISE and NOISE-CON Congress and Conference* (pp. 477–488).

Taal, C. H., Hendriks, R. C., Heusdens, R., & Jensen, J. (2011). An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*, 2125–2136.

Tan, K., & Wang, D. L. (2019a). Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 6865–6869).

Tan, K., & Wang, D. L. (2019b). Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 380–390.

Tan, L., & Jiang, J. (2001). Adaptive Volterra filters for active control of nonlinear noise processes. *IEEE Transactions on signal processing*, *49*, 1667–1676.

Tarabini, M., & Roure, A. (2008). Modeling of influencing parameters in active noise control on an enclosure wall. *Journal of Sound and Vibration*, *311*, 1325–1339.

Tobias, O. J., & Seara, R. (2005). Leaky-FXLMS algorithm: stochastic analysis for Gaussian data and secondary path modeling error. *IEEE Transactions on Speech and Audio Processing*, *13*, 1217–1230.

Tobias, O. J., & Seara, R. (2006). On the LMS algorithm with constant and variable leakage factor in a nonlinear environment. *IEEE transactions on signal processing*, *54*, 3448–3458.

Tokhi, M., & Wood, R. (1997). Active noise control using radial basis function networks. *Control Engineering Practice*, *5*, 1311–1322.

Varga, A., & Steeneken, H. J. (1993). 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*, *12*, 247–251.

Wang, D. L., & Chen, J. (2018). Supervised speech separation based on deep

Williamson, D. S., Wang, Y., & Wang, D. L. (2016). Complex ratio masking for joint enhancement of magnitude and phase. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 5220–5224).

Yang, F., Cao, Y., Wu, M., Albu, F., & Yang, J. (2018). Frequency-domain filtered-x LMS algorithms for active noise control: a review and new insights. *Applied Sciences*, *8*, 2313.

Zhang, H., & Wang, D. L. (2020). A deep learning approach to active noise control. In *Proceedings of the 2020 Conference of the International Speech Communication Association* (pp. 1141–1145).

Zhang, Q.-Z., Gan, W.-S., & Zhou, Y.-l. (2006). Adaptive recurrent fuzzy neural networks for active noise control. *Journal of Sound and Vibration*, *296*, 935–948.

Zhou, D., & DeBrunner, V. (2007). Efficient adaptive nonlinear filters for nonlinear active noise control. *IEEE Transactions on Circuits and Systems I: Regular Papers*, *54*, 669–681.

Zhou, Y.-L., Zhang, Q.-Z., Li, X.-D., & Gan, W.-S. (2005). Analysis and DSP implementation of an ANC system using a filtered-error neural network. *Journal of Sound and Vibration*, *285*, 1–25.

Zhu, Q., Qiu, X., & Burnett, I. (2020). An acoustic modelling based remote error sensing approach for quiet zone generation in a noisy environment. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 8424–8428).

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: