Neural Cascade Architecture for Multi-Channel Acoustic Echo Suppression

Hao Zhang, *Member, IEEE*, and DeLiang Wang, *Fellow, IEEE*

***Abstract*—Traditional acoustic echo cancellation (AEC) works by identifying an acoustic impulse response using adaptive algorithms. This paper proposes a neural cascade architecture for joint acoustic echo and noise suppression to address both single-channel and multi-channel AEC (MCAEC) problems. The proposed cascade architecture consists of two modules. A convolutional recurrent network (CRN) is employed in the first module for complex spectral mapping. Its output is fed as an additional input to the second module, where a long short-term memory network (LSTM) is utilized for magnitude mask estimation. The entire architecture is trained in an end-to-end manner with the two modules optimized jointly using a single loss function. The final output is generated using the enhanced phase and magnitude obtained from the first and the second module, respectively. The cascade architecture enables the proposed method to obtain robust magnitude estimation as well as phase enhancement. The proposed method is investigated under different AEC setups. We find that the deep learning based approach avoids the no-uniqueness problem in traditional MCAEC. For MCAEC setups with multiple microphones, combining deep MCAEC with supervised beamforming further improves the system performance. Evaluation results show that the proposed approach effectively suppresses acoustic echo and noise while preserving speech quality, and consistently outperforms related methods under different setups.**

***Index Terms*—Acoustic echo cancellation, deep learning, neural cascade architecture, multi-channel AEC, nonlinear distortions.**

I. Introduction

ACOUSTIC echo arises when a loudspeaker and a microphone are coupled in a communication system such that the microphone picks up the desired near-end speech plus the loudspeaker signal. If not properly handled, a user at the far end of the system hears his or her own voice in the loudspeaker signal delayed by the round trip time of the system (i.e. an echo), mixed with the target speech signal from the near end. The acoustic echo is one of the most annoying problems in speech telecommunication such as teleconferencing, hands-free telephony, and mobile communication.

Conventionally, acoustic echo is removed by adaptively identifying an acoustic impulse response between the loudspeaker and the microphone using a finite impulse response (FIR) filter [1]. Many algorithms, such as normalized least mean square (NLMS), affine projection, and recursive least squares algorithms, have been proposed [1]–[4]. The performance of these algorithms depends on how well their parameters control the speed of convergence while keeping misalignment in check. Especially when near-end and far-end speakers both talk (i.e. double-talk), convergence rates have to compromise between the two. Moreover, nonlinear distortions may be introduced to the recordings due to the poor quality of electronic devices such as amplifiers and loudspeakers [5]. Traditional AEC algorithms are linear and fundamentally cannot handle nonlinear distortions [5], [6].

To enhance user experience, modern hands-free communication devices are commonly equipped with multiple loudspeakers and multiple microphones. The availability of additional devices makes it necessary to design multi-channel acoustic echo cancellation (MCAEC), which presents additional challenges and opportunities compared to single-channel AEC. Although conceptually similar, MCAEC is fundamentally different from single-channel AEC and a straightforward extension of single-channel AEC does not result in satisfactory performance. First, the number of echo paths to be modeled is increased and the convergence of the individual filter could depend on the performance of other filters. Therefore, it is crucial to use well-designed step-size control for proper convergence of all the filters. Second is the well-known non-uniqueness problem [3], [7], which arises because multi-channel far-end signals are typically highly correlated. As a result, the echo paths cannot be determined uniquely, impacting the convergence of adaptive techniques [7]. Many methods have been proposed to circumvent this problem [3], [8], [9], among which coherence reduction methods are most commonly used. Such methods, however, inevitably degrade the perceptual audio quality of the signals reproduced by loudspeakers, and a compromise must be made between convergence and audio quality [10].

In MCAEC with multiple microphones, microphone array speech enhancement techniques such as beamforming (BF) can be combined with AEC for efficient reduction of noise and acoustic echoes [11]. The most straightforward combination applies them in sequence, i.e., applying single-channel AEC to each microphone signal before beamforming or applying single-channel AEC to the output of a beamformer [11]. In general, the former outperforms the latter since the beamformer

Manuscript received 20 December 2021; revised 4 May 2022 and 3 July 2022; accepted 14 July 2022. Date of publication 19 July 2022; date of current version 27 July 2022. This work was supported in part by NSF under Grant ECCS-1808932 and in part by the Ohio Supercomputer Center. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Lei Xie.

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

DeLiang 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.2022.3192104

in the latter introduces time variations to the echo path and affects AEC convergence [12], [13]. Other algorithms employ relative echo transfer functions [14], [15] or joint optimization strategies [16], [17] to improve the MCAEC performance. However, these strategies tend to introduce convergence issues and effective combinations of AEC and beamforming are yet to be found [10].

Deep learning has been utilized for addressing AEC problems due to its capacity in modeling complex nonlinear relations. Birkett and Goubran [18] use a cascaded time-delay neural network to model the nonlinearity of the acoustic channel. Lee *et al.* [19] use a neural network as a residual echo suppressor to remove the nonlinear echo components. Zhang and Wang [20] formulate AEC as supervised speech separation and suppress echo by extracting the near-end speech from a microphone recording. Carbajal *et al.* [21] utilize a multi-input neural network to estimate phase-sensitive masks. Early studies focus on magnitude enhancement and use magnitude masks as the training targets [19]–[22]. Recently, complex-domain estimation is employed for phase enhancement to improve the quality of estimated near-end speech. A convolutional recurrent neural network (CRN) is introduced to perform complex spectral mapping for echo suppression [23]. AEC methods using complex-valued neural networks for phase-aware enhancement are studied in [24], [25]. AEC Challenges [26], [27] show that various deep learning architectures can be utilized to address AEC problems. There is a trend of combining traditional and deep AEC methods as multi-stage systems where a traditional algorithm is utilized in the first stage for initial echo removal and deep learning is used in the second stage for residual echo suppression [24], [28], [29]. Moreover, multiple neural networks have been combined to perform joint echo and noise suppression [30]–[32].

Motivated by a recently introduced neural cascade architecture for speech enhancement [33], we propose a neural cascade architecture (NCA) for joint acoustic echo and noise suppression. The proposed cascade architecture consists of two modules where a CRN is used in the first module for complex spectral mapping. The estimated magnitude is then used as an additional input in the second module for magnitude mask estimation, which allows for progressive enhancement of the target speech. Different from previous multi-stage studies that employ sequential training steps with separate loss functions, the proposed cascade architecture is trained in an end-to-end manner using a single loss function. Training the two modules simultaneously using the single loss function allows for the correction of estimation errors of the first module. Finally, the estimated magnitude from the second module, together with the enhanced phase from the first module, is used to generate time-domain near-end speech. Hence the cascade architecture leverages the advantages of the two modules and is expected to obtain robust magnitude as well as phase estimation.

We further extend the neural cascade architecture to AEC with multi-loudspeakers and multi-microphones. Instead of estimating acoustic echo paths, our deep MCAEC works by directly estimating near-end speech, which intrinsically avoids the non-uniqueness problem in traditional MCAEC algorithms. Although there are multiple acoustic paths in MCAEC setups, the proposed approach naturally addresses the problem through model training using a deep learning model, rather than training a separate AEC model for each echo path. In addition, our deep MCAEC can be combined with supervised beamforming with a microphone array to further improve echo suppression performance.

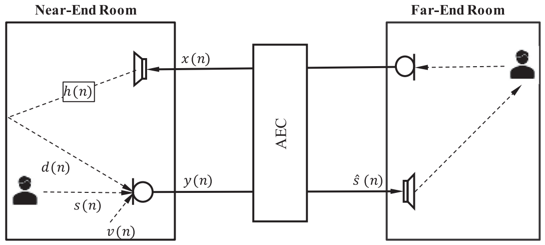


Fig. 1. Typical single-channel acoustic echo cancellation system.

Compared to the preliminary version [34], this paper introduces a different network structure, conducts more extensive evaluations, and investigates the performance of the proposed method under different AEC setups. Comparisons with other deep learning based methods are provided to show the effectiveness of the proposed method.

The remainder of this paper is organized as follows. Section II describes the problem and introduces magnitude and complex domain estimation. Section III presents the proposed neural cascade architecture. Its multi-channel version is introduced in Section IV. Evaluation metrics and experiment settings are given in Section V. In Section VI, we present evaluation and comparison results. Section VII concludes this paper.

II. Problem Formulation

A. Signal Model

1) Single Channel AEC: Fig. 1 illustrates a typical single-channel AEC system. The far-end signal $x(n)$ is played through the loudspeaker in the near end and reaches the microphone via the acoustic echo path $h(n)$. The microphone signal $y(n)$ is a mixture of echo $d(n)$, near-end speech $s(n)$, and background noise $v(n)$:

$$y(n) = d(n) + s(n) + v(n) \tag{1}$$

where n indexes a time sample and the echo signal is generated by convolving the loudspeaker signal (or the transformed version of $x(n)$) with a room impulse response (RIR) between loudspeaker and microphone, $h(n)$.

Traditional AEC algorithms achieve echo removal by adaptively estimating the acoustic echo path given $y(n)$ and $x(n)$ as inputs. Then the estimated echo signal $\hat{d}(n)$ is subtracted from the microphone signal to get the system output. Residual echo suppression and post-filtering are usually used to further suppress residual echo and noise.

2) Multi-Channel AEC: The demand for MCAEC has increased in recent years with the rise of hands-free devices and teleconferencing systems. To provide realistic audio effects, multiple loudspeakers are used, to achieve enhanced sound quality, microphone arrays are used. This transforms a single-channel AEC structure to a more complex MCAEC structure. For a general MCAEC setup with L loudspeakers and M microphones, the signal picked up by microphone m is represented as:

$$y_m(n) = \sum_{l=1}^{L} d_{lm}(n) + s_m(n) + v_m(n) \tag{2}$$

where $m = 1, 2, \ldots, M$, and $d_{lm}(n)$ denotes the echo signal from loudspeaker l to microphone m at the near end.

There are two major differences between MCAEC and single-channel AEC. First, the system topology is different and the number of echo paths increases from one to ML. Ideally, there should be an adaptive filter for each echo path. Second, when exposed to highly correlated input channels, the convergence of adaptive techniques is affected and the echo paths cannot be determined uniquely (non-uniqueness problem).

B. Deep AEC

The ultimate goal of AEC is to transmit only near-end speech to the far end. From the speech separation point of view, AEC can be naturally considered as a supervised speech separation problem where near-end speech is the target source to be separated from the microphone recording. Therefore, instead of estimating the acoustic echo path, a deep learning based approach works by directly estimating the near-end speech from the microphone signal with the accessible far-end speech as an additional input. Many deep AEC methods have been proposed in the literature and these methods can be generally categorized into two groups: magnitude mask estimation and complex-domain estimation. Magnitude mask estimation focuses on the magnitude spectrogram of the target signal and estimates masks, such as the ideal ratio mask (IRM) and phase sensitive mask (PSM), in the time-frequency (T-F) domain. The other group uses complex-domain estimation to achieve both magnitude and phase enhancement.

Echo and noise suppression performance depends on the accuracy of estimated magnitudes while the quality of separated near-end speech also depends on phase information. Magnitude mask estimation based methods utilize ideal ratio masks as the training targets. The value range of these masks is usually bounded to $[0, 1]$, which facilitates mask estimation and leads to better echo and noise suppression. Complex-domain methods jointly estimate magnitude and phase, resulting in improvement in speech quality. However, the value range of the real and imaginary targets used in complex-domain estimation methods, either complex ratio mask [35] or complex spectral mapping [36], is unbounded. Although techniques have been proposed to bound the output range [37], it is still harder to achieve a robust magnitude estimate compared to magnitude mask estimation methods.

III. Neural Cascade Architecture

We start by introducing the proposed approach for single-channel AEC. The proposed neural cascade architecture consists of a complex module and a magnitude mask module, as shown in Fig. 2. The main idea of this design is to leverage the advantages of complex-domain estimation and magnitude mask estimation so as to obtain phase enhancement as well as robust magnitude estimation. We have also explored other cascading options such as using magnitude mask estimation before complex-domain estimation. The architecture presented in this paper achieves the best overall performance.

A. Complex Module

The complex module employs a CRN for complex spectral mapping. The CRN takes the real and imaginary spectrograms of microphone and far-end signals $[Y_{(r)}(t,f), Y_{(i)}(t,f), X_{(r)}(t,f), X_{(i)}(t,f)]$ as inputs to predict the real and imaginary spectrograms of near-end speech $[\hat{S}'_{(r)}(t,f), \hat{S}'_{(i)}(t,f)]$, where $Y(t,f)$, $X(t,f)$, and $\hat{S}'(t,f)$ are the short-time Fourier transform (STFT) of microphone signal, far-end signal, and estimated near-end speech obtained from the first module within the T-F unit at time t and frequency f, respectively, and sbuscripts (r) and (i) denote the real and imaginary parts of the corresponding signals. The enhanced magnitude and phase are then calculated, respectively, as:

$$|\hat{S}'(t,f)| = \sqrt{\hat{S}'^2_{(r)}(t,f) + \hat{S}'^2_{(i)}(t,f)} \tag{3}$$

$$\theta_{\hat{S}'}(t,f) = \arctan\left(\hat{S}'_{(i)}(t,f) / \hat{S}'_{(r)}(t,f)\right) \tag{4}$$

The CRN has an encoder-decoder architecture with a two-layer grouped LSTM in the bottleneck to model temporal dependencies. The encoder and decoder comprise five convolutional layers and five deconvolutional layers, respectively, as illustrated in Fig. 2. A detailed description of the CRN architecture is provided in [36] except that our CRN has four input channels.

B. Magnitude Mask Module

The estimated $|\hat{S}'(t,f)|$, together with $|Y(t,f)|$ and $|X(t,f)|$ are fed to the magnitude mask module to predict a T-F mask $M(t,f)$ using an LSTM network. The estimated magnitude spectrogram is obtained as:

$$|\hat{S}(t,f)| = M(t,f) \odot |Y(t,f)| \tag{5}$$

where $\odot$ denotes element-wise multiplication.

The final output, near-end speech $\hat{s}(n)$, is generated by feeding the estimated magnitude $|\hat{S}(t,f)|$ and the enhanced phase from the complex module $\theta_{\hat{S}'}(t,f)$ to inverse short time Fourier transform (iSTFT):

$$\hat{s}(n) = \text{iSTFT}\left(|\hat{S}(t,f)|, \theta_{\hat{S}'}(t,f)\right) \tag{6}$$

The LSTM has four hidden layers with 300 units in each layer. The output is fully connected. In our implementation, we bound the value range of the mask output between $[0, 1]$, and use the sigmoid function as the activation function in the output layer.

C. Loss Functions and Model Training

The training objective of the cascade architecture consists of two parts, corresponding to the complex and magnitude mask

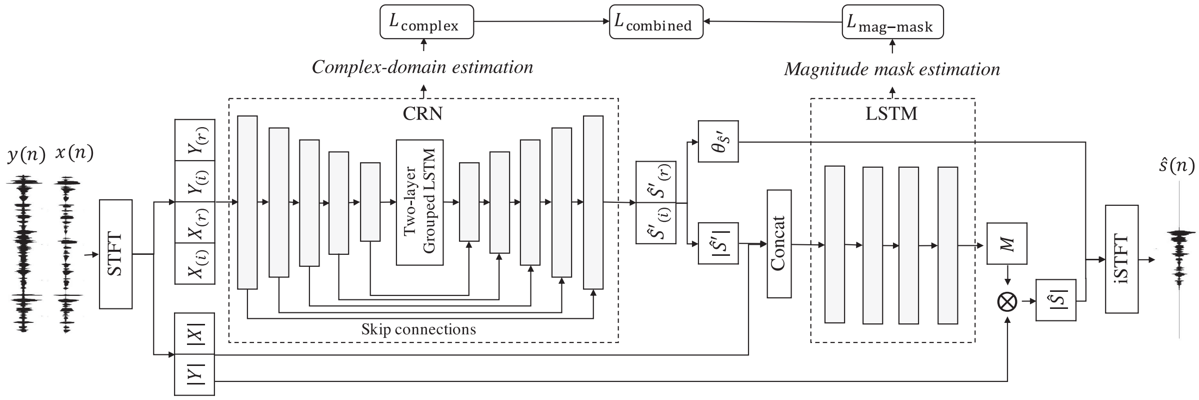


Fig. 2. Diagram of the proposed neural cascade architecture for joint echo and noise suppression. The first module employs a CRN for complex spectral mapping, the output is concatenated with original inputs and fed to an LSTM to predict T-F masks. Subscripts (r) and (i) denote real and imaginary spectrograms of signals, respectively, $\theta_{\hat{S}'}$ denotes the phase of $\hat{S}'$, and $|\cdot|$ denotes magnitude.

modules. Following [38], we define the first loss L_{complex} as the real, imaginary, and magnitude differences between $S'(t,f)$ and $S(t,f)$:

$$L_{\text{complex}} = \frac{1}{TF}\sum_{t,f}(|S'_{(r)}(t,f) - S_{(r)}(t,f)|^2 + |S'_{(i)}(t,f) - S_{(i)}(t,f)|^2 + ||S'(t,f)| - |S(t,f)||^2) \quad (7)$$

where T and F denote the number of time frames and frequency bins, respectively. The second loss corresponding to the magnitude mask module is given below:

$$L_{\text{mag-mask}} = \frac{1}{TF}\sum_{t,f}\left(|\hat{S}(t,f) - |S(t,f)|\right)^2 \quad (8)$$

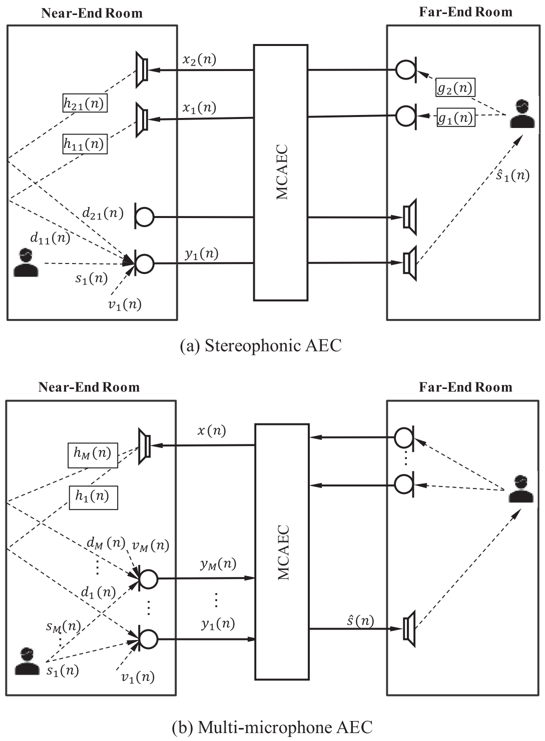
Rather than undergoing multiple sequential training stages with separate loss functions, we propose to combine L_{complex} and $L_{\text{mag-mask}}$ and train the cascade architecture only once with the following single loss function:

$$L_{\text{combined}} = \lambda L_{\text{complex}} + (1-\lambda)L_{\text{mag-mask}} \quad (9)$$

where λ is a coefficient for combining the two losses, and set to $2/3$ based on the performance on validation data.

IV. Neural Cascade Architecture for MCAEC

This section expands the proposed neural cascade architecture to address MCAEC problems. Broadly speaking, we investigate two different MCAEC setups, stereophonic AEC and AEC setup with a microphone array. The configurations of these two MCAEC setups are provided in Fig. 3.



(a) Stereophonic AEC

(b) Multi-microphone AEC

Fig. 3. Two multi-channel acoustic echo cancellation setups: (a) stereophonic AEC, and (b) AEC setup with multiple microphones.

A. Stereophonic AEC

As shown in Fig. 3(a), a stereophonic AEC (SAEC) system is composed of two microphones and two loudspeakers, respectively. The far-end signals $x_1(n)$ and $x_2(n)$ are generated by convolving a common source with two RIRs, $g_1(n)$ and $g_2(n)$, and then transmitted to two loudspeakers in the near end with its own microphones. The signal picked up by microphone m at the near end consists of two echo signals $d_{1m}(n)$, $d_{2m}(n)$, near-end speech $s_m(n)$, and background noise $v_m(n)$:

$$y_m(n) = \sum_{l=1}^{2} d_{lm}(n) + s_m(n) + v_m(n),\ m = 1,2 \quad (10)$$

Instead of estimating all the acoustic echo paths, NCA-based SAEC directly estimates the near-end speech. Therefore, the proposed SAEC method avoids the non-uniqueness problem and there is no need to decorrelate the stereophonic signals of d_{1m} and d_{2m}.

Two training strategies are proposed to train the SAEC models. The first strategy trains a single-input and single-output (SISO) network and estimates each target $s_m(n)$ individually given $y_m(n)$, $x_1(n)$ and $x_2(n)$ as inputs. The training signals used in this strategy are sampled from the two microphones with m randomly selected from $\{1, 2\}$. Through this, the model is exposed to all the microphones during training and a model trained this way can be used to achieve echo suppression for both microphones in the SAEC system. The second strategy trains a multi-input and multi-output (MIMO) network that predicts the target speech at all microphones jointly. Specifically, it utilizes the two microphone signals and two far-end signals as inputs to simultaneously estimate the near-end speech signals received by the microphones. SAEC models trained using these two strategies are found to produce similar results while the first strategy requires smaller input and output dimensions, which results in less training time, especially under MCAEC setup with multiple microphones.

B. MCAEC With Multiple Microphones

Considering a multi-microphone AEC (MMAEC) setup with M microphones and one loudspeaker, as shown in Fig. 3(b). The signal received at microphone m can be represented as:

$$y_m(n) = d_m(n) + s_m(n) + v_m(n),\ m = 1, 2, \cdots M \tag{11}$$

where $d_m(n)$ denotes the echo.

Different MMAEC methods have been discussed in [11]–[13]. In this paper, we apply AEC separately to each microphone signal before beamforming [11]. The SISO training strategy is employed for model training. During training, we use $y_m(n)$ and $x(n)$ as inputs and set the corresponding near-end speech $s_m(n)$ as the training target.

Once the model is trained, the estimated speech is used for complex spectral mapping (CSM) based minimum variance distortion-less response (MVDR) beamforming [38]. Designating the first microphone as the reference microphone, the MVDR beamformer is given as:

$$\hat{\boldsymbol{w}}(t, f) = \frac{\hat{\Phi}_N^{-1}(t,f)\hat{\boldsymbol{c}}(t,f)}{\hat{\boldsymbol{c}}(t,f)^H\hat{\Phi}_N^{-1}(t,f)\hat{\boldsymbol{c}}(t,f)} \tag{12}$$

where superscript H denotes conjugate transpose, $\hat{\Phi}_N(t, f)$ is the estimated covariance matrix of interference (acoustic echo and background noise), $\hat{\boldsymbol{c}}(t, f)$ is the estimated steering vector, which is estimated as the principal eigenvector of the estimated speech covariance matrix $\hat{\Phi}_S(t, f)$ [39]. The estimated covariance matrices of speech and interference are obtained as follows

$$\hat{\Phi}_S(t, f) = \frac{1}{t}\sum_{t'=1}^{t} \hat{\boldsymbol{S}}(t,' f)\hat{\boldsymbol{S}}^H(t,' f) \tag{13}$$

$$\hat{\Phi}_N(t, f) = \frac{1}{t}\sum_{t'=1}^{t} \hat{\boldsymbol{N}}(t,' f)\hat{\boldsymbol{N}}^H(t,' f) \tag{14}$$

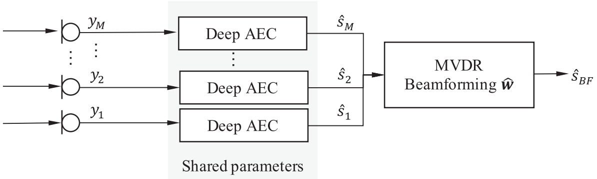


Fig. 4. Deep learning for multi-microphone AEC.

where $\hat{\boldsymbol{S}}(t,' f)$ is the STFT representation of all the CSM estimated speech signals at time t' and frequency f, and $\hat{\boldsymbol{N}}(t,' f) = \boldsymbol{Y}(t,' f) - \hat{\boldsymbol{S}}(t,' f)$ is the estimated interference [38].

The MVDR beamformer is applied to array signals and the enhanced results are calculated as:

$$\hat{S}_{BF}(t, f) = \hat{\boldsymbol{w}}^H(t, f)\hat{\boldsymbol{S}}(t, f) \tag{15}$$

The beamformer performs spatial filtering to maintain signals from the desired direction while suppressing interferences from other directions. The overall structure of the deep MMAEC is shown in Fig. 4. Note that the covariance matrices and beamformer are updated each time frame using only past information, hence the MVDR is a casual beamformer. To ensure that the calculated covariance matrices are nonsingular, the first 3 frames of each signal are grouped together in our implementation to calculate the beamformers for these frames. Compared to the non-causal version in a previous study [34], the causal MVDR beamformer achieves similar PESQ values and slightly lower ERLE values.

Besides SAEC and MMAEC, the proposed method can handle the most general situation with arbitrary numbers of microphones and loudspeakers. For such a setup, via SISO or MIMO training we can estimate the near-end speech at each microphone, and the estimated signals can be further enhanced using CSM-based MVDR beamforming before feeding to the corresponding loudspeakers at the far end.

V. Experimental Setup

A. Experiment Setup

The TIMIT dataset [40] is used to perform experiments in AEC situations with double-talk, background noise, and nonlinear distortions. TIMIT contains 6300 sentences from 630 speakers. From these speakers, we randomly choose 100 pairs of speakers (40 pairs of male-female, 30 male-male, and 30 female-female) as near-end and far-end speakers. Out of the ten utterances of each speaker, seven are randomly chosen to create training mixtures, and the remaining three to create test mixtures. To be specific, the three chosen utterances from a far-end speaker are concatenated to generate a far-end signal. A randomly chosen utterance from a near-end speaker is extended to the same length as that of the far-end signal by zero padding at the beginning and the end, with the number of leading zeros randomly chosen between zero and the maximum number that needs no zero padding at the end. By doing this, the starting position of near-end speech is randomized in each mixture.

RIRs are simulated using the image method [41]. To investigate RIR generalization, we simulate 20 rooms of different sizes $a \times b \times c$ m^3 for training, where $a \in \{4, 6, 8, 10\}, b \in \{5, 7, 9, 11, 13\}$, and $c = 3$. For single-channel AEC, we generate ten pairs of random positions in each room with a fixed microphone-loudspeaker distance (1 m) and a fixed microphone to near-end speaker distance (0.5 m) to simulate RIRs for loudspeaker and near-end speaker. For stereophonic AEC setup, the two microphones and the two loudspeakers are positioned at $(a/2, b/2 + 0.05, c/2)$ m, $(a/2, b/2 - 0.05, c/2)$ m, $(a/2, b/2 + 0.6, c/2 + 0.5)$ m, and $(a/2, b/2 - 0.6, c/2 + 0.5)$ m, respectively. The near-end speaker is placed at 20 random positions in each room with 1 m apart from the center of the microphones. The MMAEC setup consists of a uniform linear array with four microphones and one loudspeaker. The center of the microphone array is positioned at the center of the room with a 4 cm inter-microphone distance. Twenty pairs of positions are simulated randomly for the loudspeaker and the near-end speaker in each room, and the distance from the loudspeaker and the near-end speaker to the center of the array are set to 0.6 m and 1 m, respectively. The reverberation time (T_{60}) is randomly selected from $\{0.2, 0.3, 0.4, 0.5, 0.6\}$ s. For testing, we simulate three untrained rooms of size $3 \times 4 \times 3$ m, $5 \times 6 \times 3$ m, $11 \times 14 \times 3$ m, set T_{60} to 0.35 s, and generate ten pairs of RIRs by placing the near-end speaker at 10 random positions in each room. The corresponding RIRs are denoted as RIR1, RIR2, and RIR3, respectively.

We consider different nonlinear distortions in our experiments [19], [23], [42], [43]. In [19] the nonlinear distortions introduced by a power amplifier and a loudspeaker are simulated in the following steps. First, hard clipping [44] is applied to each far-end signal to simulate the characteristic of a power amplifier:

$$x_{\texttt{hard}}(n) = \begin{cases} -x_{\texttt{max}} & x(n) < -x_{\texttt{max}} \\ x(n) & |x(n)| \leq x_{\texttt{max}} \\ x_{\texttt{max}} & x(n) > x_{\texttt{max}} \end{cases} \tag{16}$$

where $x_{\texttt{max}}$ is set to 0.8 as the maximum amplitude of $|x(n)|$. Then a sigmoidal nonlinearity [45] is applied to the clipped signal to simulate an asymmetric loudspeaker distortion:

$$x_{\texttt{NL}}(n) = \gamma \left(2/\left(1 + \exp\left(-a \cdot b(n)\right)\right) - 1 \right) \tag{17}$$

where $b(n) = 1.5 \times x_{\texttt{hard}}(n) - 0.3 \times x^2_{\texttt{hard}}(n)$. The gain γ is set to 4, and the slope a is set to 4 if $b(n) > 0$ and 0.5 otherwise. These parameter values are taken from [45].

Another commonly used nonlinear distortion for loudspeaker is the saturation type simulated using the scaled error function (SEF) [42]:

$$f_{\text{SEF}}(x) = \int_0^x e^{-\frac{z^2}{2\eta^2}} dz \tag{18}$$

where x is the input to the loudspeaker, and η^2 represents the strength of nonlinearity. The SEF becomes linear as η^2 tends to infinity and becomes a hard limiter as it tends to zero. 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).

We use 10000 noises from a sound effect library (http://www.sound-ideas.com) to create training mixtures (see [46]) for single-channel AEC setup. Operational room noise (oproom), speech shaped noise (SSN) from NOISEX-92 dataset [47], babble noise from an Auditec CD (http://www.auditec.com), and white noise are used for creating test mixtures. For multi-channel AEC setup, the babble noise from the NOISEX-92 dataset is used as the background noise and the algorithm proposed in [48] is employed to make the noise diffuse. The diffuse babble noise is then split into two parts, the first 80% of which is used for training and the remaining 20% is used for testing.

We create 20000 training and 300 test mixtures. Each training mixture is created by first passing a far-end signal through the nonlinear model to generate a loudspeaker signal. The loudspeaker signal is then convolved with a randomly chosen RIR from the training RIRs for the loudspeaker to generate an echo signal. A randomly chosen near-end utterance is convolved with an RIR for the near-end speaker and then mixed with the echo at a signal-to-echo ratio (SER) randomly chosen from $\{-6, -3, 0, 3, 6\}$ dB. Finally, a noise of the same length is added to the mixture at a signal-to-noise ratio (SNR) randomly chosen from $\{8, 10, 12, 14\}$ dB. The SER and SNR, which are evaluated during double-talk periods, are defined as:

$$\text{SER} = 10 \log_{10} \left[\sum_n s^2(n) / \sum_n d^2(n) \right] \tag{19}$$

$$\text{SNR} = 10 \log_{10} \left[\sum_n s^2(n) / \sum_n v^2(n) \right] \tag{20}$$

Test mixtures are created similarly but using different utterances, noises, RIRs, SERs and SNRs.

B. Comparison Methods

We compare with four deep learning methods for single-channel AEC setup. The LSTM baseline is a causal version of the method proposed in [20]. It achieves echo suppression by estimating a magnitude mask and using the phase of microphone signal. The CRN method [23] achieves echo suppression through complex spectral mapping and estimates the real and imaginary spectrograms of near-end speech jointly. The multi-input residual echo suppression (MI-RES) method is a two-stage system that combines an adaptive algorithm with a neural network [21]. The neural network in MI-RES is used to estimate a phase sensitive mask of near-end speech with the estimated echo as additional input. The LFM-NFM [28] is a cascaded deep AEC method that consists of a linear-filtering model (LFM) and a nonlinear-filtering model (NLM), where the LFM removes the linear part of the echo and the NLM handles the nonlinear residual echo.

For stereophonic AEC setup, we compare with a stereophonic joint-optimized normalized least mean square algorithm (SJONLMS) and two deep learning methods. SJONLMS is a stereophonic version of joint-optimized normalized least mean square algorithm (JONLMS) [49] equipped with a coherence reduction technique described in [50]. The global step size of JONLMS is adjusted iteratively based on joint optimization with respect to system misalignment. Therefore, it is robust to double-talk and achieves faster convergence and lower misalignment compared to the NLMS algorithms that use a constant step size. Post-filtering (PF) [51] is employed to further suppress

TABLE I
AEC RESULTS IN THE PRESENCE OF DOUBLE-TALK, WHITE NOISE, NONLINEAR DISTORTIONS AND UNTRAINED RIRS (RIR1) WITH 10 DB SNR

	3.5 dB		0 dB		-3.5 dB	
	ERLE	PESQ	ERLE	PESQ	ERLE	PESQ
Unprocessed	0	1.96 ± 0.19	0	1.80 ± 0.20	0	1.60 ± 0.26
MI-RES [21]	33.26 ± 2.41	2.28 ± 0.19	33.56 ± 2.50	2.22 ± 0.18	35.04 ± 2.71	2.13 ± 0.19
LSTM [20]	44.67 ± 6.78	2.38 ± 0.19	45.80 ± 7.04	2.24 ± 0.20	46.73 ± 7.63	2.08 ± 0.24
CRN [23]	35.64 ± 2.44	2.64 ± 0.18	36.87 ± 2.66	2.51 ± 0.18	37.50 ± 2.54	2.33 ± 0.21
CRN-L	35.15 ± 1.34	**2.69 ± 0.17**	35.93 ± 1.69	**2.54 ± 0.18**	36.12 ± 2.19	2.36 ± 0.20
LFM-NFM [28]	38.44 ± 4.46	2.59 ± 0.21	41.42 ± 4.27	2.45 ± 0.22	44.32 ± 4.11	2.26 ± 0.23
Proposed	**53.37 ± 6.59**	2.68 ± 0.15	**53.97 ± 6.71**	**2.54 ± 0.16**	**53.58 ± 7.15**	**2.38 ± 0.19**

residual noise and echo (SJONLMS-PF). The parameters of SJONLMS and PF are given in [49]–[51]. CRN-complex is a CSM based MCAEC method [34]. CRN-mask focuses on SAEC and employs CRN with gated recurrent unit for magnitude mask estimation [52].

For the MCAEC setup with multiple microphones, we employ single-channel JONLMS [49] for each microphone as a baseline and then combine the outputs with the ideal MVDR beamformer (JONLMS-IBF). The ideal MVDR beamformer (IBF) is calculated by inserting the true speech and interference signals ($\boldsymbol{S}(t,f)$ and $\boldsymbol{N}(t,f)$) into (12), (13), and (14). The CRN-complex based MCAEC method proposed in [34] is employed as another comparison method.

Signals are sampled at 16 kHz, and windowed into 20-ms frames with 10-ms frame shift. Then a 320-point STFT is applied to each frame to produce a spectrogram. All the deep networks are trained for 30 epochs with a learning rate of 0.001. We apply utterance level normalization to the input mixtures in our experiments. The normalization is implemented by dividing the signal by the root mean square power of the microphone signal.

C. Evaluation Metrics

AEC performance is evaluated in terms of ERLE [4] for single-talk periods and perceptual evaluation of speech quality (PESQ) [53] for double-talk periods. Evaluation results are presented as mean ± standard deviation (std).

ERLE is the most commonly used metric for AEC and it is defined as:

$$\mathrm{ERLE} = 10\log_{10}\left[\sum_n y^2(n) / \sum_n \hat{s}^2(n)\right] \qquad (21)$$

This variant of ERLE is widely used for evaluating AEC systems in the presence of background noise [20], [28], [29], [54]. It considers both echo and noise and is calculated as the ratio of the input energy to the output energy. PESQ is a widely used quality metric for speech enhancement, and its values range from -0.5 to 4.5. For both metrics, a higher score indicates better performance.

TABLE II
AEC RESULTS USING DIFFERENT CASCADING OPTIONS UNDER 3.5 DB SER, 10 DB SNR, AND BABBLE NOISE

3.5 dB SER	ERLE	PESQ
Unprocessed	-	2.00 ± 0.20
CRN-Complex + CRN-Complex	37.36 ± 1.98	2.59 ± 0.20
LSTM-Mask + LSTM-Mask	43.85 ± 6.38	2.47 ± 0.18
LSTM-Mask + CRN-Complex	37.46 ± 2.34	2.60 ± 0.19
CRN (phase) + LSTM (magnitude)	44.52 ± 6.66	2.50 ± 0.17
Proposed	**52.34 ± 6.95**	**2.67 ± 0.15**
Proposed (1^{st} module only)	35.21 ± 1.75	2.65 ± 0.14
Proposed (2^{nd} module only)	51.83 ± 6.79	2.66 ± 0.15

TABLE III
AEC RESULTS USING THE SAME NETWORK STRUCTURE BUT DIFFERENT TRAINING STRATEGIES UNDER RIR1, 3.5 DB SER, 10 DB SNR AND WHITE NOISE

3.5 dB SER	ERLE	PESQ
Unprocessed	0	1.96 ± 0.19
Proposed	**53.37 ± 6.59**	**2.68 ± 0.15**
Magnitude masking on S'_m	51.80 ± 7.65	2.65 ± 0.18
Multi-stage sequential training	49.47 ± 6.51	2.59 ± 0.16
Optimizing $L_{\text{mag-mask}}$ only	47.00± 5.55	2.52 ± 0.19

TABLE IV
AEC RESULTS UNDER UNTRAINED SPEAKERS, RIRS, AND ECHO PATH CHANGES WITH 3.5 DB SER,10 DB SNR, AND WHITE NOISE

	ERLE	PESQ	
	Proposed	Unprocessed	Proposed
Untrained speaker	53.57 ± 6.30	1.95 ± 0.23	2.69 ± 0.19
RIR2	54.99 ± 5.27	1.95 ± 0.18	2.75 ± 0.15
RIR3	56.67 ± 3.84	1.92 ± 0.14	2.79 ± 0.15
Echo path change	53.48 ± 6.47	1.96 ± 0.18	2.68 ± 0.17

TABLE V
SAEC RESULTS FOR PROPOSED METHOD USING DIFFERENT TRAINING STRATEGIES UNDER 3.5 DB SER, 10 DB SNR, RIR2, AND LINEAR DISTORTIONS

	Mic1		Mic2	
	ERLE	PESQ	ERLE	PESQ
Unprocessed	0	2.11 ± 0.19	0	2.07 ± 0.19
SISO	56.45 ± 6.93	2.80 ± 0.15	58.28 ± 5.71	2.83 ± 0.15
MIMO	56.63 ± 3.87	2.80 ± 0.13	57.07 ± 3.86	2.79 ± 0.14

VI. EXPERIMENTAL RESULTS

A. Single-Channel AEC

1) Evaluation and Comparison Results: We first evaluate the proposed method under the single-channel AEC setup and compare it with other deep learning methods. The evaluation results in situations with double-talk, untrained background noise, untrained RIRs, and nonlinear distortions are presented in Table I. These methods can all suppress echo and the proposed method consistently outperforms baseline methods in terms of ERLE and PESQ. The proposed method, which can be regarded

TABLE VI
SAEC RESULTS OF PROPOSED AND BASELINE METHODS IN THE PRESENCE OF DOUBLE-TALK, BACKGROUND NOISE WITH 3.5 DB SER, 10 DB SNR, AND LINEAR DISTORTIONS

Mic1	RIR1		RIR2		RIR3	
	ERLE	PESQ	ERLE	PESQ	ERLE	PESQ
Unprocessed	0	2.08 ± 0.19	0	2.07 ± 0.18	0	2.11 ± 0.20
SJONLMS	7.37 ± 1.86	2.38 ± 0.17	7.80 ± 1.95	2.46 ± 0.13	7.75 ± 1.72	2.46 ± 0.13
SJONLMS-PF	20.10 ± 2.94	2.49 ± 0.20	21.30 ± 2.56	2.55 ± 0.12	20.64 ± 2.04	2.57 ± 0.10
CRN-complex [34]	29.10 ± 6.36	2.57 ± 0.18	35.77 ± 2.32	2.78 ± 0.16	37.32 ± 0.94	**2.90 ± 0.15**
CRN-mask [52]	39.10 ± 5.78	2.55 ± 0.16	42.75 ± 3.46	2.72 ± 0.16	48.12 ± 3.83	2.80 ± 0.16
Proposed	**50.01 ± 10.64**	**2.60 ± 0.17**	**57.27 ± 7.09**	**2.80 ± 0.16**	**59.58 ± 4.49**	**2.90 ± 0.15**

TABLE VII
SAEC RESULTS IN THE PRESENCE OF DOUBLE-TALK, BACKGROUND NOISE WITH RIR2, 10 DB SNR, 3.5 DB SER, AND NONLINEAR DISTORTIONS

Mic1	$\eta^2 = 0.1$		$\eta^2 = 0.5$	
	ERLE	PESQ	ERLE	PESQ
Unprocessed	0	2.07 ± 0.19	0	2.06 ± 0.19
CRN-complex [34]	35.78 ± 2.39	2.78 ± 0.16	35. 78 ± 2.64	2.78 ± 0.17
CRN-mask [52]	42.64 ± 3.36	2.71 ± 0.16	42.78 ± 3.49	2.72 ± 0.17
Proposed	**56.91 ± 6.51**	**2.79 ± 0.15**	**57.26 ± 6.42**	**2.79 ± 0.16**

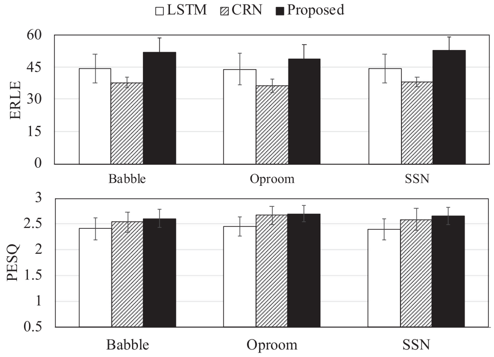


Fig. 5. ERLE and PESQ values in the presence of different untrained noises, with SER = 3.5 dB, SNR = 10 dB, RIR1, and nonlinear distortions.

as a combination of the LSTM and CRN methods, further improves ERLE of LSTM by 8.76 in situations with 3.5 dB SER. Compared with the CRN-based CSM, the proposed method improves ERLE significantly while achieving comparable and slightly better speech quality. Moreover, we increase the model size of the CRN baseline to a comparable level to the size of the proposed method (see Fig. 7 later). The resulting larger model, denoted as CRN-L, achieves a little higher PESQ and slightly lower ERLE scores compared to the CRN baseline. Compared to the proposed method, CRN-L achieves comparable PESQ results but lower ERLE values. Fig. 5 shows the results of the proposed method tested under different untrained noises, which indicate that the proposed method generalizes well to different untrained noise.

Table II shows the AEC results for cascade architectures with different combinations of the complex module (CRN-Complex) and magnitude mask module (LSTM-Mask). Besides different cascade architectures, three other rows are provided for further comparisons. Among the rows, CRN (phase) + LSTM (magnitude) stands for the method that trains the CRN and LSTM model separately, and then generates the final output using the estimated phase and magnitude, respectively, from the CRN and the LSTM model. Proposed (1st module only) and Proposed (2nd module only) refer to the enhanced results obtained from the first and second module of the proposed method alone. The table shows that the proposed neural cascade architecture achieves the best overall performance.

2) Training Strategies: This part evaluates the proposed cascade architecture trained using different strategies. The results under RIR1, 3.5 dB SER, 10 dB SNR, and white noise are given in Table III. There are two reasonable masking strategies for the magnitude mask module in the proposed architecture: applying the estimated magnitude mask to microphone signal Y_m or the estimated near-end speech S'_m. The models trained using these two strategies are comparable while the proposed method achieves slightly better performance. This is because the estimated S'_m contains distortions, and applying a mask to S'_m could further distort speech components. The fourth row of the table shows the results of the cascade architecture trained sequentially using separate loss functions. The last row gives the results of the proposed architecture trained by only optimizing the loss function at the final output, $L_{\text{mag-mask}}$. Comparisons show that the proposed method, which is trained in an end-to-end manner using a single combined loss function, outperforms the alternative training strategies. This illustrates that the strong performance of the proposed method is due to not only the neural network structure, but also the combined loss function and the training strategy.

3) Robustness Test: We further test the proposed method in situations with untrained speakers, untrained RIRs (RIR2 and RIR3), and echo path changes to show its robustness. To create test mixtures with untrained speakers, we randomly select 10 pairs of untrained speakers from the 430 remaining TIMIT speakers and create 100 test mixtures using RIR1. The test mixtures under untrained rooms are generated using RIR2 and RIR3. The echo path change is simulated by randomly selecting two pairs of RIRs from RIR1 and switching between them every 1.5 seconds for generating each test mixture. The SER level is set to 3.5 dB and white noise is added to the mixture at an SNR level of 10 dB for all the test sets. The results given in Table IV indicate the strong robustness of the proposed method.

B. Stereophonic AEC

This part evaluates the performance of the proposed method under the stereophonic AEC setup. We first compare the performance of the proposed method trained using two different strategies described in Section IV-A, namely SISO and MIMO.

TABLE VIII
MMAEC RESULTS FOR PROPOSED AND BASELINE METHODS IN THE PRESENCE OF DOUBLE-TALK, BACKGROUND NOISE WITH RIR2, 3.5 DB SER, AND 10 DB SNR

		$\eta^2 = \infty$ (Linear)		$\eta^2 = 0.1$		$\eta^2 = 0.5$	
		ERLE	PESQ	ERLE	PESQ	ERLE	PESQ
Unprocessed		0	2.04 ± 0.19	0	2.03 ± 0.20	0	2.04 ± 0.20
JONLMS		7.75 ± 1.85	2.45 ± 0.14	-	-	-	-
JONLMS-IBF		16.83 ± 1.86	2.64 ± 0.15	-	-	-	-
CRN-complex [34]	$\hat{s}$	35.14 ± 1.01	2.80 ± 0.19	35.18 ± 0.98	2.80 ± 0.20	35.18 ± 1.05	2.81 ± 0.20
	y_{BF}	5.02 ± 1.13	2.33 ± 0.19	5.10 ± 1.20	2.33 ± 0.19	5.13 ± 1.19	2.33 ± 0.19
	$\hat{s}_{BF}$	35.90 ± 1.29	2.85 ± 0.23	35.85 ± 1.33	2.84 ± 0.24	35.88 ± 1.35	**2.85 ± 0.23**
Proposed	$\hat{s}$	55.41 ± 6.06	2.81 ± 0.17	56.24 ± 5.36	2.81 ± 0.17	**55.61 ± 5.73**	2.82 ± 0.17
	y_{BF}	7.41 ± 1.99	2.33 ± 0.19	7.43 ± 1.54	2.31 ± 0.19	7.16 ± 1.52	2.32 ± 0.19
	$\hat{s}_{BF}$	**55.67 ± 6.14**	**2.86 ± 0.18**	**56.47 ± 5.57**	**2.85 ± 0.18**	55.56 ± 6.34	**2.85 ± 0.18**

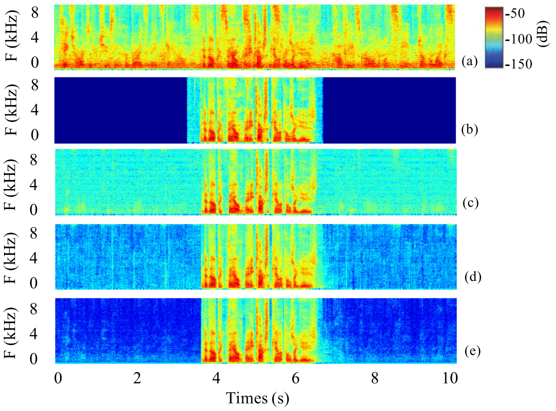


Fig. 6. Spectrograms of a test sample under SAEC setup with 3.5 dB SER, 10 dB SNR (babble noise), RIR2, and nonlinear distortions: (a) microphone signal, (b) near-end target speech, and outputs of (c) CRN-complex, (d) CRN-mask, and (e) Proposed.

The results tested on both microphones (Mic1 and Mic2) are given in Table V. It is seen that the models trained using these two strategies obtain similar results. Considering that the SISO strategy requires less training time, we utilize it for model training in the following experiments.

Table VI compares the results obtained using different SAEC methods in situations with double-talk, diffused babble noise, and linear distortions. Since the two microphones are handled in the same way and achieve similar results, only the results tested at Mic1 are presented in this table. The proposed method consistently outperforms the other comparison methods in terms of ERLE and PESQ and the performance generalizes well to untrained RIRs.

The results of the methods in situations with nonlinear distortions are provided in Table VII. All these deep learning methods can handle nonlinear distortions and the proposed method achieves the best results. Fig. 6 presents the spectrograms of a test sample in situations with double-talk, background noise, and nonlinear distortions. It is evident that the proposed method achieves the best echo suppression and has the least residual echo and noise in the enhanced speech.

C. Multi-Microphone AEC

The performance of the proposed method under MMAEC setup is evaluated in this subsection. Three results are provided for each deep learning method, where $\hat{s}$ is the enhanced signal obtained at the reference microphone, y_{BF} and $\hat{s}_{BF}$ are, respectively, the time-domain beamformed microphone signal and beamformed enhanced signal introduced in Section IV-B.

TABLE IX
SAEC AND MMAEC RESULTS OF PROPOSED METHOD UNDER UNTRAINED AND MOVING SPEAKERS WITH 3.5 DB SER,10 DB SNR, RIR2, AND NONLINEAR DISTORTIONS ($\eta^2 = 0.1$)

		Untrained speakers		Moving speakers	
SAEC		ERLE	PESQ	ERLE	PESQ
Unprocessed		0	2.10 ± 0.22	0	2.11 ± 0.20
Proposed		57.32 ± 5.72	2.80 ± 0.17	55.36	2.79 ± 0.14
MMAEC		ERLE	PESQ	ERLE	PESQ
Unprocessed		0	2.06 ± 0.18	0	2.06 ± 0.23
Proposed	$\hat{s}$	56.37 ± 5.08	2.75 ± 0.16	55.04 ± 5.85	2.77 ± 0.17
	y_{BF}	6.88 ± 2.14	2.32 ± 0.17	7.89 ± 1.88	2.36 ± 0.19
	$\hat{s}_{BF}$	56.67 ± 5.39	2.86 ± 0.15	55.87 ± 5.97	2.87 ± 0.15

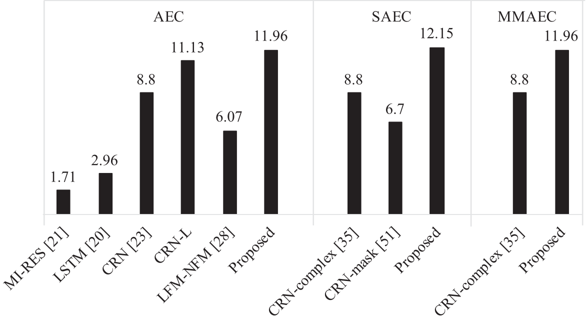


Fig. 7. Numbers of trainable parameters (in million) for different models.

The comparison results with both linear and nonlinear distortions are summarized in Table VIII. In general, the NCA-based MMAEC outperforms traditional and CRN-complex methods. All deep learning methods can suppress most of the echo and noise from microphone recordings, as seen from the $\hat{s}$ results. Combining with MVDR beamformer ($\hat{s}_{BF}$) further improves the overall performance in almost all the cases.

To test robustness, Table IX shows the behavior of the proposed SAEC and MMAEC methods in situations with untrained speakers and moving speakers. The test signals for untrained speakers are created by randomly selecting 10 pairs of new speakers from the TIMIT dataset. The test signals for moving speakers are generated by changing the position of a near-end speaker at the middle point of an utterance. To simulate this, we randomly select two RIRs generated for the near-end speaker from RIR2 and convolve them with the first half and second half of the utterance, respectively. The results in this table demonstrate the strong robustness of the proposed methods.

Model sizes (the number of trainable parameters within a model) of the baselines and the proposed model are presented in Fig. 7. The MI-RES has the smallest number of parameters. The proposed neural cascade architecture has about 11.96 million parameters for the single-channel AEC and MMAEC cases and 12.15 million parameters for the SAEC case. Our model achieves strong performance with reasonable model sizes.

VII. Conclusion

This paper introduces a neural cascade architecture to address the multi-channel AEC problem, where single-channel AEC becomes a special case. The proposed method cascades complex spectral mapping and magnitude mask estimation in order to leverage their advantages to achieve phase and magnitude enhancement jointly. The cascade architecture is trained using a single loss function in an end-to-end manner. The final output is obtained using the enhanced magnitude from the magnitude mask module and the enhanced phase from the complex module. Experimental results demonstrate that the proposed method outperforms related deep AEC methods and generalizes well to untrained scenarios. Moreover, the proposed method overcomes the limitations of traditional MCAEC methods and produces superior ERLE and PESQ scores. Combining deep MCAEC and CSM-based beamforming further improves the system performance. Future work will explore the performance of deep MCAEC using real-recorded signals and address practical issues such as computational complexity.

Acknowledgment

The authors would like to thank K. Tan for his assistance in using the CRN code, and H. Wang for helpful discussion regarding [33].

References

[1] J. Benesty, T. Gänsler, D. Morgan, M. Sondhi, and S. Gay, *Advances in Network and Acoustic Echo Cancellation*. Berlin/Heidelberg, Germany: Springer, 2001.

[2] E. Hänsler and G. Schmidt, *Acoustic Echo and Noise Control: A Practical Approach*, vol. 40. Hoboken, NJ, USA: Wiley, 2005.

[3] J. Benesty, C. Paleologu, T. Gänsler, and S. Ciochină, *A Perspective on Stereophonic Acoustic Echo Cancellation*. Berlin/Heidelberg, Germany: Springer, 2011, vol. 4.

[4] G. Enzner, H. Buchner, A. Favrot, and F. Kuech, "Acoustic echo control," in *Academic Press Library in Signal Processing*, vol. 4. Amsterdam, The Netherlands: Elsevier, 2014, pp. 807–877.

[5] A. Birkett and R. A. Goubran, "Limitations of handsfree acoustic echo cancellers due to nonlinear loudspeaker distortion and enclosure vibration effects," in *Proc. Workshop Appl. Signal Process. Audio Acoust.*, 1995, pp. 103–106.

[6] M. I. Mossi, N. W. Evans, and C. Beaugeant, "An assessment of linear adaptive filter performance with nonlinear distortions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2010, pp. 313–316.

[7] M. M. Sondhi, D. R. Morgan, and J. L. Hall, "Stereophonic acoustic echo cancellation-an overview of the fundamental problem," *IEEE Signal Process. Lett.*, vol. 2, no. 8, pp. 148–151, Aug. 1995.

[8] M. Schneider and W. Kellermann, "Multichannel acoustic echo cancellation in the wave domain with increased robustness to nonuniqueness," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 3, pp. 518–529, Mar. 2016.

[9] J. Franzen and T. Fingscheidt, "An efficient residual echo suppression for multi-channel acoustic echo cancellation based on the frequency-domain adaptive Kalman filter," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 226–230.

[10] M. and L. Valero, "Acoustic echo reduction for multiple loudspeakers and microphones: Complexity reduction and convergence enhancement," Ph.D. dissertation, Friedrich-Alexander-Univ Erlangen-Nürnberg, Erlangen, Germany, 2019.

[11] W. Kellermann, "Strategies for combining acoustic echo cancellation and adaptive beamforming microphone arrays," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 1997, vol. 1, pp. 219–222.

[12] W. Herbordt and W. Kellermann, "Limits for generalized sidelobe cancellers with embedded acoustic echo cancellation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2001, vol. 5, pp. 3241–3244.

[13] S. Doclo, M. Moonen, and E. De Clippel, "Combined acoustic echo and noise reduction using GSVD-based optimal filtering," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2000, vol. 2, pp. II1061–II1064.

[14] G. Reuven, S. Gannot, and I. Cohen, "Joint noise reduction and acoustic echo cancellation using the transfer-function generalized sidelobe canceller," *Speech Commun.*, vol. 49, pp. 623–635, 2007.

[15] M. L. Valero and E. A. Habets, "Multi-microphone acoustic echo cancellation using relative echo transfer functions," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2017, pp. 229–233.

[16] W. Herbordt, W. Kellermann, and S. Nakamura, "Joint optimization of LCMV beamforming and acoustic echo cancellation," in *Proc. 12th Eur. Signal Process. Conf.*, 2004, pp. 2003–2006.

[17] W. Herbordt and W. Kellermann, "GSAEC–acoustic echo cancellation embedded into the generalized sidelobe canceller," in *Proc. 10th Eur. Signal Process. Conf.*, 2000, pp. 1–4.

[18] A. N. Birkett and R. A. Goubran, "Acoustic echo cancellation using NLMS-neural network structures," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 1995, pp. 3035–3038.

[19] C. M. Lee, J. W. Shin, and N. S. Kim, "DNN-based residual echo suppression," in *Proc. INTERSPEECH*, 2015.

[20] H. Zhang and D. L. Wang, "Deep learning for acoustic echo cancellation in noisy and double-talk scenarios," in *Proc. INTERSPEECH*, 2018, pp. 3239–3243.

[21] G. Carbajal, R. Serizel, E. Vincent, and E. Humbert, "Multiple-input neural network-based residual echo suppression," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2018, pp. 231–235.

[22] H. Seo, M. Lee, and J.-H. Chang, "Integrated acoustic echo and background noise suppression based on stacked deep neural networks," *Appl. Acoust.*, vol. 133, pp. 194–201, 2018.

[23] H. Zhang, K. Tan, and D. L. Wang, "Deep learning for joint acoustic echo and noise cancellation with nonlinear distortions," in *Proc. INTERSPEECH*, 2019, pp. 4255–4259.

[24] M. M. Halimeh, T. Haubner, A. Briegleb, A. Schmidt, and W. Kellermann, "Combining adaptive filtering and complex-valued deep postfiltering for acoustic echo cancellation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 121–125.

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

[26] K. Sridhar *et al.*, "ICASSP 2021 acoustic echo cancellation challenge: Datasets, testing framework, and results," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 151–155.

[27] R. Cutler *et al.*, "INTERSPEECH 2021 acoustic echo cancellation challenge," in *Proc. INTERSPEECH*, 2021, pp. 4748–4752.

[28] C. Zhang and X. Zhang, "A robust and cascaded acoustic echo cancellation based on deep learning," in *Proc. INTERSPEECH*, 2020, pp. 3940–3944.

[29] J. M. Valin, S. Tenneti, K. Helwani, U. Isik, and A. Krishnaswamy, "Low-complexity, real-time joint neural echo control and speech enhancement based on percepnet," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 7133–7137.

[30] X. Shu *et al.*, "Joint echo cancellation and noise suppression based on cascaded magnitude and complex mask estimation," 2021, *arXiv:2107.09298*.

[31] N. L. Westhausen and B. T. Meyer, "Acoustic echo cancellation with the dual-signal transformation LSTM network," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 7138–7142.

[32] R. Peng, L. Cheng, C. Zheng, and X. Li, "ICASSP 2021 acoustic echo cancellation challenge: Integrated adaptive echo cancellation with time alignment and deep learning-based residual echo plus noise suppression," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 146–150.

[33] H. Wang and D. L. Wang, "Neural cascade architecture with triple-domain loss for speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 734–743, 2022.

[34] H. Zhang and D. Wang, "A deep learning approach to multi-channel and multi-microphone acoustic echo cancellation," in *Proc. INTERSPEECH*, 2021, pp. 1139–1143.

[35] D. S. Williamson, Y. Wang, and D. L. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 3, pp. 483–492, Mar. 2016.

[36] K. Tan and D. L. Wang, "Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2019, pp. 6865–6869.

[37] H. S. Choi, J. H. Kim, J. Huh, A. Kim, J. W. Ha, and K. Lee, "Phase-Aware speech enhancement with deep complex U-Net," in *Proc. Int. Conf. Learn. Representations*, 2018.

[38] Z. Q. Wang and D. L. Wang, "Deep learning based target cancellation for speech dereverberation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 941–950, 2020.

[39] J. Benesty, I. Cohen, and J. Chen, *Fundamentals of Signal Enhancement and Array Signal Processing*. Hoboken, NJ, USA: Wiley, 2018.

[40] L. F. Lamel, R. H. Kassel, and S. Seneff, "Speech database development: Design and analysis of the acoustic-phonetic corpus," in *Proc. Speech Input/Output Assessment Speech Databases*, 1989.

[41] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoustical Soc. Amer.*, vol. 65, pp. 943–950, 1979.

[42] F. Agerkvist, "Modelling loudspeaker non-linearities," in *Proc. AES Conf.: 32nd Int. Conf.: DSP Loudspeakers*, Audio Engineering Society, 2007.

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

[44] S. Malik and G. Enzner, "State-space frequency-domain adaptive filtering for nonlinear acoustic echo cancellation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 7, pp. 2065–2079, Sep. 2012.

[45] D. Comminiello, M. Scarpiniti, L. A. Azpicueta-Ruiz, J. Arenas-Garcia, and A. Uncini, "Functional link adaptive filters for nonlinear acoustic echo cancellation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1502–1512, Jul. 2013.

[46] J. Chen, Y. Wang, S. Yoho, D. Wang, and E. Healy, "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," *J. Acoustical Soc. Amer.*, vol. 139, pp. 2604–2612, 2016.

[47] 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.

[48] E. A. Habets, I. Cohen, and S. Gannot, "Generating nonstationary multisensor signals under a spatial coherence constraint," *J. Acoustical Soc. Amer.*, vol. 124, pp. 2911–2917, 2008.

[49] C. Paleologu, S. Ciochină, J. Benesty, and S. L. Grant, "An overview on optimized NLMS algorithms for acoustic echo cancellation," *Eurasip J. Adv.*, vol. 2015, 2015, Art. no. 97.

[50] M. Djendi, "An efficient stabilized fast newton adaptive filtering algorithm for stereophonic acoustic echo cancellation SAEC," *Comput. Elect. Eng.*, vol. 38, pp. 938–952, 2012.

[51] F. Ykhlef and H. Ykhlef, "A post-filter for acoustic echo cancellation in frequency domain," in *Proc. 2nd World Conf. Complex Syst.*, 2014, pp. 446–450.

[52] L. Cheng, R. Peng, A. Li, C. Zheng, and X. Li, "Deep learning-based stereophonic acoustic echo suppression without decorrelation," *J. Acoustical Soc. Amer.*, vol. 150, pp. 816–829, 2021.

[53] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2001, vol. 2, pp. 749–752.

[54] K. Nathwani, "Joint acoustic echo and noise cancellation using spectral domain Kalman filtering in double-talk scenario," in *Proc. 16th Int. Workshop Acoust. Signal Enhancement*, 2018, pp. 1–330.