



# LCSM: A Lightweight Complex Spectral Mapping Framework for Stereophonic Acoustic Echo Cancellation

Chenggang Zhang, Jinjiang Liu, Xueliang Zhang

Department of Computer Science, Inner Mongolia University, China  
cgzhang@mail.imu.edu.cn, jetliu1994@foxmail.com, cszxl@imu.edu.cn

## Abstract

The traditional adaptive algorithms will face the non-uniqueness problem when dealing with stereophonic acoustic echo cancellation (SAEC). In this paper, we first propose an efficient multi-input and multi-output (MIMO) scheme based on deep learning to filter out echoes from all microphone signals at once. Then, we employ a lightweight complex spectral mapping framework (LCSM) for end-to-end SAEC without decorrelation preprocessing to the loudspeaker signals. Inplace convolution and channel-wise spatial modeling are utilized to ensure the near-end signal information is preserved. Finally, a cross-domain loss function is designed for better generalization capability. Experiments are evaluated on a variety of untrained conditions and results demonstrate that the LCSM significantly outperforms previous methods. Moreover, the proposed causal framework only has 0.55 million parameters, much less than the similar deep learning-based methods, which is important for the resource-limited devices.

**Index Terms:** end-to-end, lightweight, stereophonic acoustic echo cancellation

## 1. Introduction

Stereophonic teleconferencing systems utilize spatial information to improve sound quality and realism compared with the single channel setting (i.e., one microphone and one loudspeaker). Traditional stereophonic acoustic echo cancellation (SAEC) methods usually remove the undesired echoes by estimating the acoustic echo paths between the stereophonic loudspeakers and microphones [1, 2]. However, the non-uniqueness problem is caused by the strong linear correlation between the echo signals that are played out via loudspeakers. Furthermore, the echoes are not only auto-correlated but also cross-correlated, making the SAEC ill-posed and unstable [3].

In the past decades, a number of interchannel decorrelation strategies have been proposed to mitigate the problem, e.g., addition of uncorrelated noise [4], nonlinear preprocessing [5], time-varying decorrelation [6], and non-intrusive method [7] etc. Lately, stereophonic acoustic echo suppression (SAES) without the decorrelation procedure was proposed in [8, 9], which estimates the echo spectra from the stereo signals using the Wiener filter in the short-time Fourier transform (STFT) domain. However, the above methods are based on unrealistic prior assumptions of the signal model, which inevitably degrade sound quality and stereophonic spatial perception in practice.

More recently, deep learning-based methods which intrinsically avoid the non-uniqueness problem have been proposed for solving SAEC, and achieve impressive performance. Cheng *et al.* [10] employs the convolutional recurrent network (CRN) which takes the magnitude spectra of both far-end and microphone signals as inputs to estimate a mask for the near-end magnitude spectrum. In contrast to the mask estimation, Zhang and

Wang [11] proposed a method that learn the complex spectral mapping to directly estimate real and imaginary spectra of near-end speech, by using the complex spectrums of far-end and microphone signals as the CRN inputs. However, the typical CRN utilizes the convolution with stride operation, usually greater than one, possibly resulting in the spatial information confusion for the multi-channel hidden features representation and the independence property of each individual frequency bin is destroyed.

Recently, we proposed an inplace convolution recurrent neural network for single-channel AEC [12]. In this paper, we address the SAEC problem. Specifically, we take all the far-end and microphone signals as one network input, and output the near-end signal in each microphone at once. We also restructure the network, which discards the two-decoder structure as well as the multi-task learning module in [12], in order to make the system parameters more compact. In addition, a cross-domain loss function is designed for better generalization in the different conditions.

The remainder of this paper is organized as follows. We formulate the SAEC problem in Section 2. In Section 3, we introduce the neural network of the proposed framework. The experimental setups are described in Section 4. We demonstrate and discuss the results of the proposed method in Section 5. Section 6 concludes the paper.

## 2. Problem formulation

### 2.1. Signal model

Without loss of generality, we take the common teleconferencing system which has dual-microphone and dual-loudspeaker as example. In the near-end room, the far-end signals  $x_i(k)$  are generated by a sound source convoluted with room impulse responses (RIRs)  $g_i(k)$ , where  $k$  indexes a time sample,  $i = 1, 2$ . The  $x_i^{nl}(k)$  is a nonlinear distortion of the far-end signals  $x_i(k)$  that are played by two loudspeakers. The nonlinearity of the power amplifier/loudspeaker is modeled by a 3rd-order polynomial function [12] as follows

$$x_i^{nl}(k) = 2ax(k) + ax^2(k) + x^3(k) \quad (1)$$

where  $a = \log(\varepsilon/10) + 0.1$ , and  $\varepsilon \in [2, 5]$ .

The echo signals  $e_{ij}(k)$  are generated by  $x_i^{nl}(k)$  convoluted with  $h_{ij}(k)$  which denotes the echo path from loudspeaker  $i$  to microphone  $j$ , and  $j = 1, 2$ .

$$e_{ij}(k) = x_i^{nl}(k) * h_{ij}(k). \quad (2)$$

where  $*$  denotes convolution operation. Considering the effect of reverberation of the near-end room as well, the near-end speech can be expressed by

$$s_j(k) = p(k) * g_j(k) = s_j^{early}(k) + s_j^{late}(k). \quad (3)$$

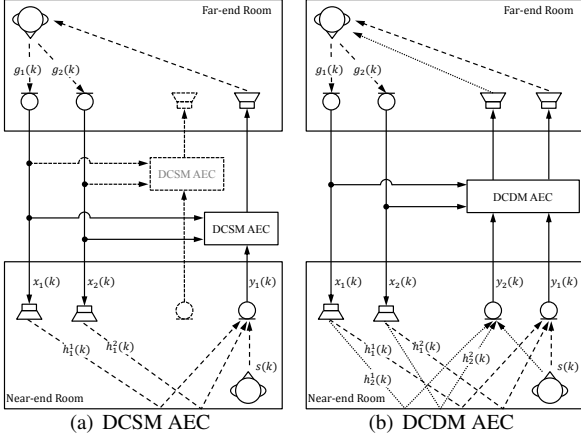


Figure 1: Diagrams of (a) Dual-Channel Single-Microphone AEC setup and (b) Dual-Channel Dual-Microphone AEC setup.

where  $p(k)$  is near-end speaker,  $s_j^{early}(k)$  denotes the early reflections of near-end signal,  $s_j^{late}(k)$  is the late reverberation reflections. Finally, the signal picked up by microphone  $j$  is mixed of echo signals  $e_{ij}(k)$  and near-end speech  $s_j(k)$  as follows

$$y_j(k) = \sum_{i=1}^2 e_{ij}(k) + s_j^{early}(k) + s_j^{late}(k). \quad (4)$$

We assume that the first microphone is always considered as the reference microphone, i. e.,  $j = 1$ . The goal of SAEC is to recover the early near-end component  $s^{early}(k)$  from  $y(k)$ , and other issues like background noise are not considered in this work.

## 2.2. Dual-Channel Single-Microphone (DCSM) AEC

Like traditional scheme, we firstly describe the SAEC problem for two loudspeakers and one microphone setting, because the other microphone needs to be handled in the same way. As illustrated in Fig. 1(a), we apply the shared DCSM to each microphone for complex spectral mapping. Specially, during the training, we stack the real and imaginary spectra of one of microphone signal ( $Y_R^j, Y_I^j$ ) and all far-end signals ( $X_R^i, X_I^i$ ) as the DCSM inputs (we omit the frame and frequency index for brevity). The complex spectra  $S_R^j$  and  $S_I^j$  of the early near-end signal  $s_j^{early}$  are the learning target, and the model can be expressed as follows

$$(S_R^j, S_I^j) = DCSM(\Upsilon | X_R^i, X_I^i, Y_R^j, Y_I^j), \quad (5)$$

$$i = 1, 2. \quad j = 1 \text{ or } 2.$$

where  $\Upsilon$  is the trainable parameter. For inference stage, the inverse STFT (iSTFT) is used to synthesize the estimated waveform of the near-end signal  $\hat{s}_j(k)$ .

## 2.3. Dual-Channel Dual-Microphone (DCDM) AEC

Considering the weakness of the DCSM needs to run two times results in high computational costs. Consequently, one solution would be designed to predict all the near-end signals simultaneously. As shown in Fig. 1(b), we proposed DCDM which consists of two loudspeakers and two microphones for joint echo cancellation. Specially, we stack all the real and imaginary spectra of microphone signals ( $Y_R^j, Y_I^j$ ) and far-end signals ( $X_R^i, X_I^i$ ) as the DCDM inputs,  $S_R^j$  and  $S_I^j$  of the near-end

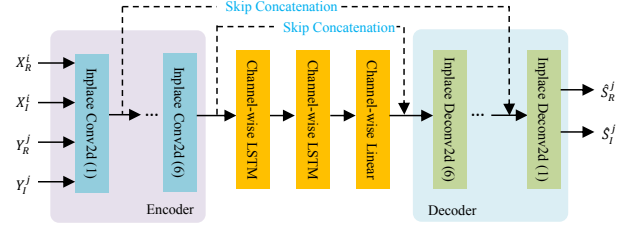


Figure 2: Network architecture of the proposed LCSM. Inplace-Conv2d and Inplace-Deconv2d layer represent the inplace convolution and deconvolution, respectively. The numbers in parenthesis indicate that encoder and decoder are symmetric structures.  $(\cdot)_R$  and  $(\cdot)_I$  are the real and imaginary spectra, respectively.

signals  $s_j^{early}$  as the training target. The model can be expressed by

$$(S_R^j, S_I^j) = DCDM(\Psi | X_R^i, X_I^i, Y_R^j, Y_I^j), i, j = 1, 2. \quad (6)$$

where  $\Psi$  is the trainable parameter. To the best of our knowledge, this is the first deep learning-based SAEC scheme that can be known as the multi-input multi-output (MIMO) system of SAEC. Since we can obtain an estimate of SAEC only once, the amount of computation is therefore dramatically reduced.

## 3. Network structure

### 3.1. Inplace convolution and channel-wise LSTM

The classic CRN [10, 11] downsample and upsample operations shrink and extend the features along the frequency dimension in the encoder and decoder, respectively. However, for multi-channel signal processing tasks, the downsampling could confuse spatial cues making LSTM hard to extract the spatial information, since the frequency information tangle with the channel dimension.

Recently, we proposed inplace convolutional recurrent neural networks, which have been used for speech enhancement and signal-channel AEC tasks [12, 13], and achieved remarkable performance. In this paper, we use a similar network structure for stereo AEC as shown in Fig. 2. Specifically, the inplace operation means that the stride of each convolution/deconvolution layer of the network is set to one, do this for keeping the number of frequency bins unchanged along the frequency dimension [13]. The size of each inplace convolution/deconvolution layer is specified in the  $[B, C, T, F]$  format which is short for  $[Batchsize, Channels, Time, Frequencies]$ . Two Channel-wise LSTM and one Linear layers are shared to extract the spatial information for each bin. The hyperparameters setting are shown in Table 1 which are given in the  $(kernelsize, strides, outchannels)$  format for convolution, and  $(hidden\_layer\_size)$  for channel-wise-LSTM and Linear layers. A further description of the inplace convolution and channel-wise LSTM is demonstrated in [12, 13]. Note that we did not use the two-decoder structure and multi-task learning module employed in [12], in order to reduce the trainable parameters.

### 3.2. Cross-domain loss function

For complex spectral mapping, the mean-squared error (MSE) loss function may be affected by the energy level of the training

Table 1: Descriptions of the hyperparameter settings of DCSM AEC. Note that for DCDM AEC, the input and output channels are set to 8 and 4, respectively.

Layer name	Input size	HyperParam.	Output size
Inplace Conv2d(1)	[B, 6, T, F]	1×5, (1, 1), 64	[B, 64, T, F]
Inplace Conv2d(2-6)	[B, 64, T, F]	1×5, (1, 1), 64	[B, 64, T, F]
Reshape	[B, 64, T, F]	-	[B×F, T, 64]
Channel-wise LSTM×2	[B×F, T, 64]	128	[B×F, T, 128]
Channel-wise Linear	[B×F, T, 128]	64	[B×F, T, 64]
Reshape	[B×F, T, 64]	-	[B, 64, T, F]
Inplace Deconv2d(6-2)	[B, 128, T, F]	1×5, (1, 1), 64	[B, 64, T, F]
Inplace Deconv2d(1)	[B, 128, T, F]	1×5, (1, 1), 64	[B, 2, T, F]

samples [14]. Recent studies [15, 16] demonstrate that combining a magnitude domain mean absolute error (MAE) loss with complex spectrum domain MAE loss is robust, because it better characterizes the distribution of normal data, and the following improved MAE (IMAE) loss function is used.

$$\mathcal{L}_{IMAE} = \|\hat{S} - |S|\|_1 + \|\hat{S}_R - S_R\|_1 + \|\hat{S}_I - S_I\|_1 \quad (7)$$

where  $\|\cdot\|_1$  is the  $L_1$  norm,  $|\hat{S}|$  and  $|S|$  denote the magnitude of the estimated signal  $\hat{s}(k)$  and target signal  $s(k)$ , respectively.

To further improve the perceptual quality of the estimated speech, we also adopted the signal-to-distortion ratio (SDR) loss which implicitly integrates phase information in time domain.

$$\mathcal{L}_{SDR} = 10 \log_{10} \frac{\|s(k)\|_2^2}{\|s(k) - \hat{s}(k)\|_2^2} \quad (8)$$

where  $\hat{s}(k)$  denotes the estimated signal, and  $\|\cdot\|_2$  is the Euclidean ( $L_2$ ) norm.

Finally, the hybrid multi-domain loss function is obtained:

$$\mathcal{L}_{CD} = \mathcal{L}_{IMAE} + \lambda \mathcal{L}_{SDR} \quad (9)$$

where the parameter  $\lambda$  is the weight factor and set to 0.1 according to our experiments. We refer to  $\mathcal{L}_{CD}$  as the cross-domain loss function, as it operates in the complex, magnitude and time-domain.

## 4. Experiments setups

### 4.1. Dataset

We use the raw speech signals from the TIMIT corpus [17] for simulating the far-end and near-end speakers as detailed description in [11, 12, 18]. Specifically, 20 different simulation rooms of size length, width and height (denote as  $l \times w \times h$ )  $m^3$  are designed, where  $l = [4, 6, 8, 10]$ ,  $w = [5, 7, 9, 11, 13]$ ,  $h = 3$ . For generating the stereo echo transmission paths, we placed two microphones in the room at positions  $(l/2, w/2 + 0.05, h/2)$  m and  $(l/2, w/2 - 0.05, h/2)$  m, respectively. One near-end speaker and the two loudspeakers are placed at 20 random locations in each room. The distance from the near-end speaker and loudspeakers to the center of the microphones are denoted as  $D_{nm}$  and  $D_{lm}$ , where  $D_{nm}$  randomly selected from  $\{1, 1.2, 1.4\}$  m and  $D_{lm}$  chosen from  $\{0.5, 0.7, 0.9\}$  m, respectively. It should be noted that the two loudspeakers we used are a vertically symmetrical structure about the center of the microphones. The room reverberation time ( $T_{60}$ ) is

randomly selected from  $\{0.2, 0.3, 0.4, 0.5, 0.6\}$  s, hence, the length of each RIR is  $T_{60} \times f_s$ , where  $f_s$  is signal sampling rate. In total, two sets of 6000 RIRs are created using the Image method [19]. In our experiments, we assume that the far-end room and near-end rooms are identical, so the RIRs for the speaker of the both near-end and far-end are shared. The microphone signal is generated by mixing a near-end signal and two echoes according to a signal-to-echo ratio (SER). In the training stage,  $SER \in [-9, 9]$  dB and is measured during the double-talk period. We create 30000 pairs of utterances as the training set and 2000 pairs of utterances as the validation set, respectively. For the testing experiment, we simulated the untrained room of size  $5 \times 6 \times 3$   $m^3$ ,  $D_{nm}$ ,  $D_{lm}$  and  $T_{60}$  are set to 0.6 m, 1.3 m and 0.35 s, respectively. Finally, we generate a test set of 100 pairs of utterances that is entirely distinct from training and validation sets.

### 4.2. Baselines and training setups

We utilize three algorithms as baselines for comparison experiments, there is Yang [8] which is a conventional method for addressing the stereo echo suppression problem, and Cheng *et al.* [10] and Zhang [11] are the recent CRN-based methods. The parameter settings are consistent with the description in the original papers.

In this study, all input signals are sampled to 16 kHz, and then divided into frames with 20 ms window length and 50% overlap, and Hamming window is used. The proposed LCSM is trained for 100 epochs by the Adam optimizer [20], and the batch size is set to 4. The initial learning rate is 1e-3, and dropped by a factor of 0.3 when validation loss does not improve for three epochs.

## 5. Results and discussions

Two general metrics are employed to measure the experimental results, there is the perceptual evaluation of speech quality (PESQ) [21] to measure the near-end speech quality during the double-talk period, and the echo return loss enhancement (ERLE) [22] to evaluate the echo reduction during the single-talk period, respectively. For the convenience of comparing the performance of DCSM AEC and DCDM AEC, our results are all the estimated near-end signals of the reference microphone.

First, we evaluate the performance of the algorithms on three common unmatched conditions. a) Real office RIR scenario selected from the Aachen Impulse Response database [23]. b) Unseen nonlinear distortion scenario which utilizes the *hard-clipping* function to simulate the hardware distortions as described in [24, 25], and the clipping threshold of the input signal is set to 0.7 in this work. c) Music echoes scenario, where far-end signals are music randomly chosen from MUSAN database [26], is also a very common echo in practice.

The average scores of ERLE and PESQ for all methods in different scenarios are demonstrated in Table 2. Among these methods, LCSM- $\mathcal{L}_{CD}$  represents LCSM which uses MSE loss (used in Zhang) instead of  $\mathcal{L}_{CD}$  for ablation studies. From the table, we can infer the following conclusions. 1) The proposed LCSM achieves the best scores from DCSM to DCDM settings in terms of ERLE and PESQ in most conditions, and the complexity is greatly reduced. 2) The LCSM yields better performance than LCSM- $\mathcal{L}_{CD}$ , especially in terms of ERLE. It implies that the L- $\mathcal{L}_{CD}$  significantly improves the generalization ability. 3) Because it is a statistical model of the signal, Yang achieves stable performance in terms of PESQ dur-

Table 2: Exhibition of the results under the real office RIR, untrained hard-nonlinearity and music echo conditions with different SERs.

	Metrics	Office RIR						Hard-nonlinearity						Music echo					
		ERLE ( $\uparrow$ )			PESQ ( $\uparrow$ )			ERLE ( $\uparrow$ )			PESQ ( $\uparrow$ )			ERLE ( $\uparrow$ )			PESQ ( $\uparrow$ )		
		SER (dB)	-5	0	5	-5	0	5	-5	0	5	-5	0	5	-5	0	5	-5	0
	Mixture	—	—	—	1.57	2.00	2.36	—	—	—	1.64	2.06	2.41	—	—	—	1.58	2.05	2.42
DCSM	Yang[8]	23.43	23.28	23.10	2.38	2.75	3.09	14.29	14.25	14.14	2.55	2.90	3.21	19.50	19.46	19.38	2.74	2.92	3.16
	Cheng <i>et al.</i> [10]	52.37	52.24	51.12	2.41	2.75	3.07	47.96	47.71	47.40	2.42	2.76	3.08	<b>46.01</b>	<b>43.50</b>	38.47	2.36	2.68	2.97
	Zhang[11]	42.20	41.56	39.78	2.54	2.93	3.28	41.44	40.74	39.06	2.51	2.91	3.26	33.58	30.59	26.60	2.44	2.74	2.95
	LCSM- $\mathcal{L}_{CD}$	44.15	44.29	43.40	2.82	3.22	3.57	45.56	45.41	44.20	2.89	3.28	3.61	37.27	35.81	32.81	2.76	3.04	3.47
	LCSM	<b>56.24</b>	<b>56.27</b>	<b>55.73</b>	<b>2.84</b>	<b>3.25</b>	<b>3.59</b>	<b>58.58</b>	<b>58.32</b>	<b>57.31</b>	<b>2.91</b>	<b>3.31</b>	<b>3.63</b>	41.71	40.78	<b>38.91</b>	<b>2.88</b>	<b>3.22</b>	<b>3.48</b>
DCDM	Cheng <i>et al.</i> [10]	50.60	50.66	50.45	2.43	2.77	3.09	47.99	47.85	47.40	2.45	2.79	3.10	47.89	44.98	41.21	2.37	2.71	3.02
	Zhang[11]	39.52	39.25	38.43	2.45	2.82	3.16	39.88	39.53	38.52	2.57	2.92	3.24	33.78	32.49	30.10	2.42	2.73	2.99
	LCSM- $\mathcal{L}_{CD}$	47.19	46.82	45.66	3.00	3.34	3.63	47.66	47.28	46.06	3.04	3.37	3.65	40.01	37.48	33.82	3.01	3.28	3.49
	LCSM	<b>63.78</b>	<b>63.12</b>	<b>61.68</b>	<b>3.02</b>	<b>3.35</b>	<b>3.64</b>	<b>64.02</b>	<b>63.50</b>	<b>62.12</b>	<b>3.05</b>	<b>3.38</b>	<b>3.66</b>	<b>57.46</b>	<b>53.43</b>	<b>48.23</b>	<b>3.06</b>	<b>3.32</b>	<b>3.52</b>

Table 3: Comparison of the proposed LCSM with other algorithms under the far-end is speech while near-end is music signal condition with different SERs.

	Metrics	Far-end(speech), Near-end(music)					
		ERLE ( $\uparrow$ )			SDR ( $\uparrow$ )		
		SER (dB)	-5	0	5	-5	0
DCSM	Yang[8]	14.45	14.37	14.21	-0.01	0.38	0.51
	Cheng <i>et al.</i> [10]	49.99	47.32	44.15	5.04	7.52	10.46
	Zhang[11]	41.87	40.03	37.13	5.53	7.75	10.06
	LCSM- $\mathcal{L}_{CD}$	46.25	44.66	42.23	8.79	11.49	13.14
	LCSM	<b>56.99</b>	<b>54.21</b>	<b>51.07</b>	<b>9.20</b>	<b>12.06</b>	<b>14.80</b>
DCDM	Cheng <i>et al.</i> [10]	49.04	46.97	43.79	5.13	7.42	10.07
	Zhang[11]	40.03	38.98	37.06	5.59	7.95	10.44
	LCSM- $\mathcal{L}_{CD}$	46.97	45.41	43.12	9.31	11.60	13.85
	LCSM	<b>59.41</b>	<b>56.56</b>	<b>53.45</b>	<b>10.22</b>	<b>12.59</b>	<b>15.04</b>

Table 4: Results of the number of trainable parameters for the deep learning-based SAEC. (unit is Million).

Methods	Parameters
Cheng <i>et al.</i> [10]	12.99 M
Zhang[11]	9.07 M
LCSM	0.55 M

ing the double-talk periods but yields lower scores in terms of ERLE which denotes the echo cancellation ability is limited. 4) Although Cheng *et al.* achieves impressive ERLE scores in all scenarios, the near-end signal quality is distorted seriously since the phase information is ignored. 5) The performance of LCSM- $\mathcal{L}_{CD}$  is greatly improved over Zhang, especially in terms of PESQ. And the results confirm that the in-place convolution operation and the temporal modeling of each frequency bin are valid, as we expected.

Further, we also conduct the condition in which the near-end is music but the far-end is speech. Comparisons of all methods are depicted in Table 3. We utilize SDR as the metric because PESQ is hard to measure the music signal quality. As shown in Table 3, the proposed LCSM has significant advantages over other methods in both DCSM and DCDM settings. Take SER = 0 dB in DCSM setting for example, the near-end music signal estimated by the LCSM improves SDR by 4.31 dB and improves ERLE by 14.18 dB over Zhang, respectively. Clearer spectra of different signals are exhibited in Fig. 3. Some audio samples of this scenario can be found in this page<sup>1</sup>.

Considering the limitation of memory resources, we also compare the number of trainable parameters for the deep

<sup>1</sup><https://chenggangzhang.github.io/LCSM-StereoAEC>

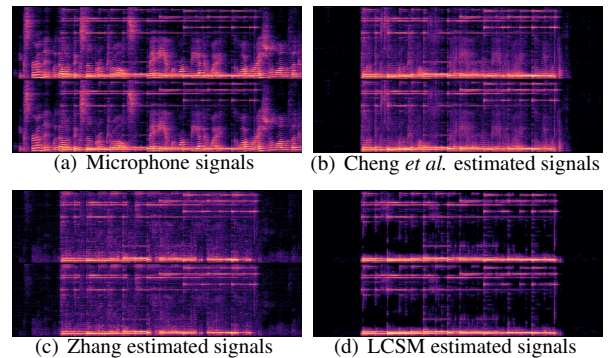


Figure 3: Presentation of spectra of different DCSM AEC methods at SER = 0 dB. Note that the top of image (a) is the first microphone signal and the bottom is the second microphone signal.

learning-based approach. Please note that the difference between DCSM and DCDM in terms of parameters is not very obvious. From Table 4, we can find that the parameters of the framework are only 0.55 M which is over 16x less than Zhang. Furthermore, the proposed LCSM is a causal system that is essential for real-time applications.

## 6. Conclusion

In this paper, a lightweight end-to-end framework is presented for solving the SAEC problem. We employ in-place convolution and channel-wise temporal modeling to keep the vital information of the near-end signal. Moreover, a cross-domain loss function is adopted for complex spectral mapping. It is shown in the performance evaluation that the proposed method has good generalization ability. In the following studies, we plan to investigate the performance of the proposed framework in real stereophonic scenarios.

## 7. Acknowledgements

The authors would like to thank Hao Zhang for his helpful discussions in the early stage. This research was supported by the China National Nature Science Foundation (No.61876214).

## 8. References

- [1] M. M. Sondhi, D. R. Morgan, and J. L. Hall, "Acoustic echo cancellation for stereophonic teleconferencing," *The Journal of the Acoustical Society of America*, vol. 94, no. 3, pp. 1826–1826, 1993.
- [2] M. Sondhi, D. Morgan, and J. Hall, "Stereophonic acoustic echo cancellation-an overview of the fundamental problem," *IEEE Signal processing letters*, vol. 2, no. 8, pp. 148–151, 1995.
- [3] J. Benesty, D. Morgan, and M. Sondhi, "A better understanding and an improved solution to the specific problems of stereophonic acoustic echo cancellation," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 156–165, 1998.
- [4] T. Gansler and P. Eneoth, "Influence of audio coding on stereophonic acoustic echo cancellation," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, vol. 6. IEEE, 1998, pp. 3649–3652.
- [5] J. Benesty, D. R. Morgan, J. L. Hall, and M. M. Sondhi, "Stereophonic acoustic echo cancellation using nonlinear transformations and comb filtering," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, vol. 6. IEEE, 1998, pp. 3673–3676.
- [6] M. Ali, "Stereophonic acoustic echo cancellation system using time-varying all-pass filtering for signal decorrelation," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, vol. 6. IEEE, 1998, pp. 3689–3692.
- [7] P. Thüne and G. Enzner, "Improved online identification of acoustic miso systems based on separated input signal components," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 413–417.
- [8] F. Yang, M. Wu, and J. Yang, "Stereophonic acoustic echo suppression based on wiener filter in the short-time fourier transform domain," *IEEE Signal Processing Letters*, vol. 19, no. 4, pp. 227–230, 2012.
- [9] C. Lee, J. Shin, and N. Kim, "Stereophonic acoustic echo suppression incorporating spectro-temporal correlations," *IEEE Signal Processing Letters*, vol. 21, no. 3, pp. 316–320, 2014.
- [10] L. Cheng, R. Peng, A. Li, C. Zheng, and X. Li, "Deep learning-based stereophonic acoustic echo suppression without decorrelation," *The Journal of the Acoustical Society of America*, vol. 150, no. 2, pp. 816–829, 2021.
- [11] H. Zhang and D. Wang, "A deep learning approach to multi-channel and multi-microphone acoustic echo cancellation," *Proc. Interspeech 2021*, pp. 1139–1143, 2021.
- [12] C. Zhang, J. Liu, and X. Zhang, "A complex spectral mapping with inplace convolution recurrent neural networks for acoustic echo cancellation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 751–755.
- [13] J. Liu and X. Zhang, "Inplace Gated Convolutional Recurrent Neural Network for Dual-Channel Speech Enhancement," in *Proc. Interspeech 2021*, 2021, pp. 1852–1856.
- [14] J. Gu, L. Cheng, X. Sun, J. Li, and Y. Yan, "Residual Echo and Noise Cancellation with Feature Attention Module and Multi-Domain Loss Function," in *Proc. Interspeech 2021*, 2021.
- [15] S. Fu, T. Hu, Y. Tsao, and X. Lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," in *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2017, pp. 1–6.
- [16] Z. Wang and D. Wang, "Deep learning based target cancellation for speech dereverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 941–950, 2020.
- [17] L. Lamel, R. Kassel, and S. Seneff, "Speech database development: Design and analysis of the acoustic-phonetic corpus," in *Speech Input/Output Assessment and Speech Databases*, 1989.
- [18] C. Zhang and X. Zhang, "A Robust and Cascaded Acoustic Echo Cancellation Based on Deep Learning," in *Proc. Interspeech 2020*, 2020, pp. 3940–3944.
- [19] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.
- [20] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [21] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "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*, vol. 2. IEEE, 2001, pp. 749–752.
- [22] G. Enzner, H. Buchner, A. Favrot, and F. Kuech, "Acoustic echo control," in *Academic Press Library in Signal Processing*. Elsevier, 2014, vol. 4, pp. 807–877.
- [23] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *International Conference on Digital Signal Processing*. IEEE, 2009, pp. 1–5.
- [24] A. Stenger and W. Kellermann, "Adaptation of a memoryless pre-processor for nonlinear acoustic echo cancelling," *Signal Processing*, vol. 80, no. 9, pp. 1747–1760, 2000.
- [25] S. Malik and G. Enzner, "State-space frequency-domain adaptive filtering for nonlinear acoustic echo cancellation," *IEEE Transactions on audio, speech, and language processing*, vol. 20, no. 7, pp. 2065–2079, 2012.
- [26] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.