Interspeech 2024
1-5 September 2024, Kos, Greece

SDAEC: Signal Decoupling for Advancing Acoustic Echo Cancellation

Fei Zhao, Jinjiang Liu, Xueliang Zhang

College of Computer Science, Inner Mongolia University, China
zhaofei@mail.imu.edu.cn, cszxl@imu.edu.cn

Abstract

In deep learning-based acoustic echo cancellation methods, neural networks implicitly learn echo paths to cancel echoes. However, under low signal-to-echo ratio conditions, the substantial energy discrepancy between the microphone signal and the reference signal impedes the network's ability, resulting in poor performance. In this study, we propose a **S**ignal **D**ecoupling-based monaural **A**coustic **E**cho **C**ancellation method called **SDAEC**. Specifically, we model the energy of the reference signal and the microphone signal to obtain an energy scaling factor. The reference signal is then multiplied by this energy scaling factor before being fed into the subsequent echo cancellation network. This approach reduces the difficulty of the subsequent echo cancellation step, thereby improving the overall cancellation performance. Experimental results demonstrate that the proposed method enhances the performance of multiple baseline models.

Index Terms: Acoustic echo cancellation, deep learning, signal decoupling

1. Introduction

Acoustic echo cancellation (AEC) remains an important research area in speech signal processing [1–3] and is designed to address acoustic coupling between loudspeakers and microphones in hands-free communication systems. The key objective is to leverage a far-end reference signal to eliminate echo in the microphone signal stemming from the far-end speaker [4]. Adaptive filtering has been the primary approach for modeling the echo path from the far-end to microphone signals [5–7]. However, a core limitation arises when dealing with nonlinear echoes. In these cases, linear methods fail to fully cancel echoes, resulting in significant residual echo. This severely degrades communication quality, especially when using lower-quality amplifiers and speakers.

In recent years, deep learning approaches have been increasingly applied to acoustic echo cancellation (AEC) in an end-to-end (E2E) manner, demonstrating strong potential. Zhang and Wang [8] proposed a deep bidirectional long short-term memory (Bi-LSTM) network to predict an ideal ratio mask for recovering the near-end signal from the combined magnitude spectra of the microphone and far-end signals. Zhang et al. [9] used convolutional recurrent networks (CRN) and LSTM to separate the near-end speech from the microphone signal. Zhang et al. [10, 11] proposed in-place convolution recurrent neural networks (ICRN), which utilize in-place convolution and channel-wise temporal modeling to preserve the near-end signal information. Zhang et al. [12] proposed MTFAA, a multi-scale time-frequency processing and streaming axial attention-based approach.

However, in scenarios with low signal-to-echo ratios (SER), the near-end signal experiences significant interference. Simultaneously, under such conditions, a substantial energy disparity often exists between the far-end reference signal and the microphone signal, posing challenges for echo cancellation. Further improvements are necessary to overcome the limitations of existing methods in addressing this situation.

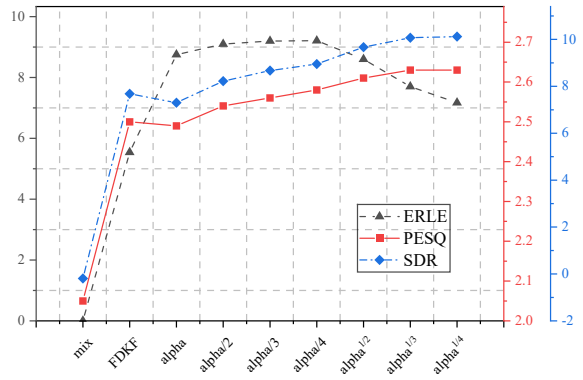


Figure 1: *FDKF and comparison of results using different alpha variants. The value alpha is calculated based on the signal-to-echo ratio (SER). alpha/n denotes dividing alpha by n, while* $alpha^{1/n}$ *represents taking the* n*th root of alpha. To reflect the substantial energy gap between the far-end reference signal and the microphone signal, we multiply the far-end reference signal by 0.1 and set the SER to -5 dB.*

To tackle this challenge, we propose a signal decoupling-based approach. When employing a neural network for AEC tasks, the network implicitly learns the echo path from the input far-end reference signal and microphone signal to achieve echo elimination. Our proposed method decouples the energy component in the echo path from the input reference signal and microphone signal, allowing for separate predictions. We denote this energy component as the energy scaling factor $alpha$. The original far-end signal is replaced by the product of the far-end reference signal with the energy scaling factor, and this product, along with the microphone signal, is used as input to the echo cancellation network. Figure 1 depicts a comparison between the unprocessed signal (mix) and the outcomes obtained by employing the traditional adaptive filtering method frequency domain Kalman filter (FDKF) [13], as well as the reference signal multiplied by the energy scaling factor $alpha$ and subsequently processed through FDKF. The variants incorporating $alpha$ showcased in the figure demonstrate further enhancement in the performance of FDKF, thereby initially validating the necessity and feasibility of our proposed method. Additionally, this observation supports our utilization

10.21437/Interspeech.2024-763

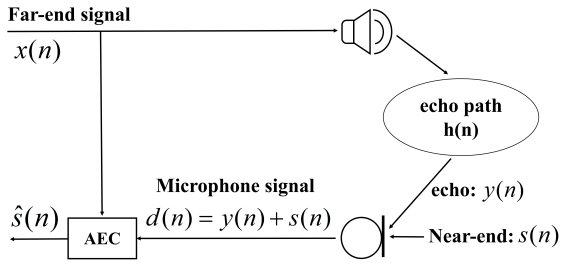


Figure 2: *Diagram of the single channel AEC.*

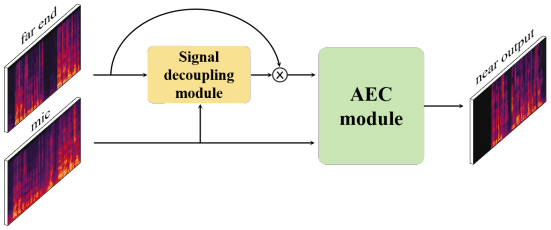


Figure 3: *SDAEC architecture overview.*

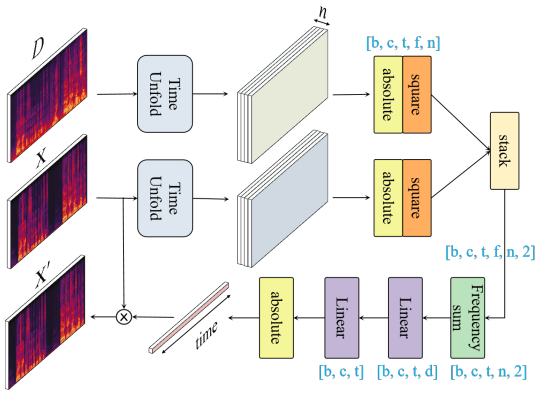


Figure 4: *Flow diagram of the signal decoupling module.*

of signal decoupling in deep learning methods.

Experimental results demonstrate that the proposed signal decoupling method not only has minimal computational requirements but can also be effectively incorporated into other end-to-end AEC systems based on deep learning, leading to performance enhancements. SDAEC script available at: https://github.com/ZhaoF-i/SDAEC

2. Problem Formulation

The diagram of the ideal single channel AEC system is illustrated in Figure 2. The microphone signal $d(n)$ is a mixture of echo signal $y(n)$ and near-end signal $s(n)$. If the environmental noise is not considered, the microphone signal in the time domain can be formulated as follows:

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

The acoustic echo $y(n)$ is the convolution of the source signal $x(n)$ with the room impulse response (RIR) [14] $h(n)$. However, in real-world scenarios, individuals often adjust the speaker volume based on the environmental conditions, causing the far-end signal $x(n)$ to be modified to $x'(n)$ through the speaker output. Simultaneously, due to the inherent limitations of the speaker itself, nonlinear distortion results in the final output of the speaker being $x'_{nl}(n)$. Then Equation 1 can be written as:

$$d(n) = x'_{nl}(n) * h(n) + s(n) \tag{2}$$

where $*$ represents the convolution operation. We reformulate Equation 2 into the time-frequency domain by applying the short-time Fourier transform (STFT) as:

$$D[t,f] = \sum_{k}^{K} H[k,f] X'_{nl}[t-k,f] + S[t,f] \tag{3}$$

where $S[t,f]$, $D[t,f]$ and $X'_{nl}[t,f]$ represent the near-end signal, microphone signal, and far-end signal played by speaker at the frame t and frequency f, respectively, and $H[k,f]$ is the echo path. Here, K stands for the number of taps.

3. Methodology

3.1. Overview

In this section, we describe the architecture of the proposed SDAEC as shown in Figure 3. The proposed approach comprises two main components: a signal decoupling module and an AEC module. First, the far-end speech signal and the microphone signal are fed into the signal decoupling module to estimate an energy scaling factor. This energy scaling factor is multiplied with the far-end speech signal to obtain a modified input. Concurrently, the microphone signal is also provided as input to the AEC module, along with the modified far-end input. The AEC module then predicts the near-end speech signal by suppressing the echo components.

3.2. Signal decoupling module

A conventional end-to-end AEC system based on deep learning can be expressed by the following Equation 4:

$$\hat{S} = F(Y, X; \Phi) \tag{4}$$

$$\hat{S} = F(Y, X'; \Phi) \tag{5}$$

where F represents the AEC network model, $\hat{S}$, Y and X represent the predicted near-end signal, microphone signal, and far-end signal in the time-frequency domain, respectively. Φ represents the network parameters.

As formulated in Equation 3, if the far-end signal played by speaker X'_{nl} can be estimated, the AEC network would only need to learn the linear echo path, significantly reducing the complexity of the learning task. However, due to the inherent nonlinearity, X'_{nl} cannot be directly obtained. Instead, we propose to exploit the energy relationship between the X and D to infer the relationship between X and X'. At this time, the AEC system can be expressed by Equation 5. This is the primary role of the signal decoupling module in our approach.

The implementation details of the signal decoupling module are shown in Figure 4. We apply a time-unfolding operation to the input far-end signal X and microphone signal D in the time-frequency domain. This operation incorporates information from the preceding (n-1) frames across the time dimension, enabling a more accurate prediction of the energy scaling factor. Subsequently, we apply the absolute value operation and the square operation separately to the time-unfolded signals, obtaining their respective energy information.

These two signals are then stacked together, and an addition operation is performed along the frequency dimension to concentrate the energy in the time dimension. The concentrated energy representation is passed through two linear layers, which map the information from n frames to a single frame and analyze the energy scaling factors from the two signals. Another absolute value operation is then applied to the output, ensuring that the resulting signal remains non-negative, thereby computing the final energy scaling factor $alpha$. Ultimately, X is multiplied by $alpha$ to obtain the modified signal X'.

3.3. Loss function

In this method, the corresponding loss function consists of multiple items. Stretched Scale-Invariant Signal-to-Noise Ratio (S-SISNR) [15] is a modified version of the Scale-Invariant Signal-to-Noise Ratio (SISNR) [16] loss function. S-SISNR is a time domain loss function that is obtained by doubling the period of SISNR. The simplified formula for S-SISNR is expressed as follows:

$$\mathcal{L}_{\text{s-sisnr}} = 10\log_{10} cot^2(\frac{\beta}{2}) = 10\log_{10}\frac{1+cos(\beta)}{1-cos(\beta)} \quad (6)$$

where β represents the angle between two vectors' ideal near-end signal s and predicted near-end signal $\hat{s}$, since it is complicated to calculate the half angle, after the derivation of the trigonometric function, it can be represented by $cos(\beta)$.

The "RI+Mag" loss criterion is adopted to recover the complex spectrum as follows:

$$\mathcal{L}_{\text{mag}} = \frac{1}{TF}\sum_{t}^{T}\sum_{f}^{F}||S(t,f)|^p - |\hat{S}(t,f)|^p|^2 \quad (7)$$

$$\mathcal{L}_{\text{RI}} = \frac{1}{TF}\sum_{t}^{T}\sum_{f}^{F}||S(t,f)|^p e^{j\theta_{S(t,f)}} - |\hat{S}(t,f)|^p e^{j\theta_{\hat{S}(t,f)}}|^2 \quad (8)$$

where p is a spectral compression factor (set to 0.5). Operator θ calculates the phase of a complex number. Then the total loss function is as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{\text{RI}} + \mathcal{L}_{\text{mag}} + \mathcal{L}_{\text{s-sisnr}} \quad (9)$$

4. Experiments

4.1. Datasets

The near-end signals are taken from the ICASSP 2023 AEC [17] challenge synthetic near-end dataset. The far-end signals and echo signals are taken from the far-end single-talk dataset in real recordings. Then echo signal is mixed with near-end speech at a SER randomly chosen from [-10, 10]dB with step 1dB. The dataset is split into 200,000 samples for the training set, 10,000 for the validation set, and 3,000 for the test set. All speech samples in the dataset have a duration of 6 seconds. Specific data generation strategies are added to the available code. It is noteworthy that the far-end single-talk dataset obtained from real recordings exhibits significant time delays. Consequently, we apply time-delay compensation to the dataset using the GCC-PHAT algorithm proposed in [18].

4.2. AEC models and training details

The proposed signal decoupling method is employed in conjunction with the following AEC models:

- **ICCRN** [19]: A neural network for monaural speech enhancement operating on the time-frequency cepstral domain. It is implemented by incorporating a cepstral frequency block into an in-place convolutional recurrent network. We change the input format of ICCRN to match the AEC task.
- **ICRN*** [10]: This model employs in-place convolution and channel-wise temporal modeling to preserve the near-end signal information. The asterisk $*$ indicates that the multi-task learning strategy employed in ICRN has been removed from this model.
- **CRN*** [9]: This model employs convolutional recurrent networks (CRN) and long short-term memory (LSTM) networks to separate the near-end speech from the microphone signal. The asterisk $*$ denotes that a convolution kernel of size (1, 3) is employed in this model.

Window length and frame shift are 20 ms and 10 ms, respectively and 320-point STFT is applied. The parameter n of the time unfold operation is set to 10. The model is optimized by Adam algorithm [20]. The initial learning rate is set to 0.001. If the validation loss does not decrease for two consecutive epochs, the learning rate is reduced by half. Training is stopped when the verification loss does not decrease for 10 consecutive epochs. The batch size is 8. It is important to note that since the echo signal originates from a real-world recording dataset, no additional loss constraints were imposed on the signal decoupling module.

4.3. Evaluation metrics

To evaluate the performance of the proposed method, we utilize the echo return loss enhancement (ERLE) [21], the perceptual evaluation of speech quality (PESQ) [3] and signal-to-distortion ratio (SDR) [22] as the metrics that measure the echo suppression for single-talk, near-end speech quality and near-end speech fidelity for double-talk periods, respectively. Higher scores indicate better performance.

5. Results

The effectiveness of the proposed signal decoupling approach is evaluated on three AEC models. As illustrated in Table 1, experiments are conducted under signal-to-echo ratios (SERs) of -10, 0, and 10 dB, encompassing simulation scenarios of double-talk, near-end single-talk, and far-end single-talk conditions.

1) For the ICCRN model:

   i) With the signal decoupling module (+SD), PESQ scores are slightly improved across all signal-to-echo ratios (SERs) in double-talk (DT), near-end single talk (ST NE), and far-end single talk (ST FE) scenarios.

   ii) The SDR metric also shows consistent improvement in all scenarios when using ICCRN+SD compared to the original ICCRN model.

   iii) In the ST FE scenario, the ERLE values are higher for ICCRN+SD across all SERs.

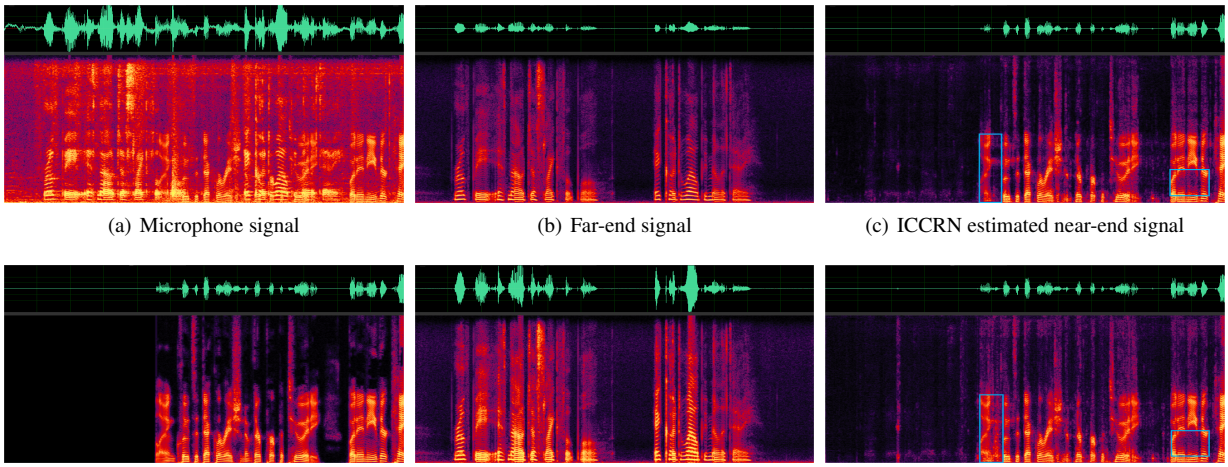
2) For the ICRN model:

   i) The addition of the signal decoupling module (+SD) leads to better PESQ scores in DT and ST NE scenarios across all SERs.

   ii) SDR values are also improved in DT and ST NE scenarios when using ICRN+SD.

Table 1: *Comparison of different AEC models combined with signal decoupling modules and the original model. SER includes -10, 0, and 10. 'DT', 'ST NE', and 'ST FE' represent double talk, near-end single talk, and far-end single talk scenarios respectively.*

SER	-10					0					10				
Scenarios	DT		ST NE		ST FE	DT		ST NE		ST FE	DT		ST NE		ST FE
Model/Metric	PESQ	SDR	PESQ	SDR	ERLE	PESQ	SDR	PESQ	SDR	ERLE	PESQ	SDR	PESQ	SDR	ERLE
Mix	1.57	-10	4.5	+∞	–	2.02	0	4.5	+∞	–	2.62	10	4.5	+∞	–
ICCRN	2.3	9.57	**4.5**	20.94	42.63	2.98	14.63	**4.5**	20.94	43.08	3.51	18.25	**4.5**	20.94	41.32
ICCRN+SD	**2.36**	**10.18**	**4.5**	**24.25**	**44.16**	**3.05**	**15.57**	**4.5**	**24.25**	**43.57**	**3.55**	**19.94**	**4.5**	**24.25**	**41.97**
ICRN	2.03	0.07	3.39	0.08	**63.12**	2.62	0.08	3.39	0.08	**52.95**	3.02	0.08	3.39	0.08	**42.83**
ICRN+SD	**2.17**	**0.14**	**3.88**	**0.17**	60.37	**2.81**	**0.17**	**3.88**	**0.17**	50.36	**3.27**	**0.18**	**3.88**	**0.17**	40.33
CRN	1.79	**3.06**	**4.42**	5.62	30.47	2.51	4.42	**4.42**	5.62	20.45	3.13	4.47	**4.42**	5.62	10.46
CRN+SD	**1.81**	2.92	4.4	**8.32**	**33.48**	**2.56**	**5.21**	4.4	**8.32**	**23.54**	**3.17**	**5.98**	4.4	**8.32**	**13.56**



(a) Microphone signal (b) Far-end signal (c) ICCRN estimated near-end signal

(d) Near-end signal (e) Far-end signal multiplied by energy scaling factor (f) ICCRN+SD estimated near-end signal

Figure 5: *Examples of waveform and spectrogram plots for different signals at SER = -10 dB.*

iii) However, in the ST FE scenario, the ERLE values are slightly lower for ICRN+SD compared to the original ICRN model.

3) For the CRN model:

i) The PESQ scores show marginal improvements when using CRN+SD in DT and ST NE scenarios across all SERs.

ii) In the ST NE scenario, the SDR values are substantially higher for CRN+SD compared to the original CRN model across all SERs.

iii) The ERLE values are also improved when using CRN+SD in the ST FE scenario.

In summary, the signal decoupling modules (+SD) consistently improve the performance of the AEC models in terms of PESQ (speech quality), SDR (signal-to-distortion ratio), and ERLE (echo suppression) metrics, especially in double-talk and near-end single talk scenarios. The improvements are more pronounced for the ICCRN and CRN models compared to the ICRN model.

Figure 5 shows a clearer waveform and spectrogram of the different signals. The green upper half of the subfigure is the waveform diagram, and the rest is the spectrogram diagram. A comparison of Figures 5(a), 5(b), and 5(e) reveals that the predicted energy scaling factor $alpha$ effectively brings the original far-end signal closer to the far-end component present in the microphone signal. Furthermore, by contrasting the blue-boxed regions in Figures 5(c) and 5(f), it becomes evident that incorporating the signal decoupling module enables more complete preservation of the near-end speech signal.

6. Conclusion

In this study, we introduce SDAEC, a novel acoustic echo cancellation method based on signal decoupling. The proposed approach estimates an energy scaling factor, denoted as $alpha$, by jointly processing the input far-end signal and the microphone signal. Subsequently, the far-end signal is scaled by the predicted $alpha$ factor and combined with the microphone signal to serve as the input to the echo cancellation model. Experimental evaluations demonstrate that the proposed SDAEC technique can be effectively integrated with various acoustic echo cancellation models, leading to further performance improvements.

Acknowledgments: This research was partly supported by the China National Nature Science Foundation (No. 61876214).

7. References

[1] M. Sondhi, "An adaptive echo canceller," *Bell System technical journal*, vol. 46, no. 3, pp. 497–511, 1967.

[2] J. Benesty, T. Gänsler, D. R. Morgan, M. M. Sondhi, S. L. Gay *et al.*, "Advances in network and acoustic echo cancellation," 2001.

[3] 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.

[4] C. Paleologu, S. Ciochina, J. Benesty, and S. L. Grant, "An overview on optimized NLMS algorithms for acoustic echo cancellation," *EURASIP J. Adv. Signal Process.*, vol. 2015, p. 97, 2015. [Online]. Available: https://doi.org/10.1186/s13634-015-0283-1

[5] S. S. Haykin, *Adaptive filter theory*. Pearson Education India, 2005.

[6] J. M. Páez-Borrallo and M. G. Otero, "On the implementation of a partitioned block frequency domain adaptive filter (PBFDAF) for long acoustic echo cancellation," *Signal Process.*, vol. 27, no. 3, pp. 301–315, 1992. [Online]. Available: https://doi.org/10.1016/0165-1684(92)90077-A

[7] N. Bershad and P. Feintuch, "Analysis of the frequency domain adaptive filter," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1658–1659, 1979.

[8] H. Zhang and D. Wang, "Deep learning for acoustic echo cancellation in noisy and double-talk scenarios," in *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018*, B. Yegnanarayana, Ed. ISCA, 2018, pp. 3239–3243. [Online]. Available: https://doi.org/10.21437/Interspeech.2018-1484

[9] H. Zhang, K. Tan, and D. Wang, "Deep learning for joint acoustic echo and noise cancellation with nonlinear distortions," in *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, G. Kubin and Z. Kacic, Eds. ISCA, 2019, pp. 4255–4259. [Online]. Available: https://doi.org/10.21437/Interspeech.2019-2651

[10] C. Zhang, J. Liu, and X. Zhang, "A complex spectral mapping with inplace convolution recurrent neural networks for acoustic echo cancellation," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*. IEEE, 2022, pp. 751–755. [Online]. Available: https://doi.org/10.1109/ICASSP43922.2022.9747459

[11] C. Zhang, J. Liu, H. Li, and X. Zhang, "Neural multi-channel and multi-microphone acoustic echo cancellation," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 2181–2192, 2023. [Online]. Available: https://doi.org/10.1109/TASLP.2023.3282103

[12] G. Zhang, L. Yu, C. Wang, and J. Wei, "Multi-scale temporal frequency convolutional network with axial attention for speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*. IEEE, 2022, pp. 9122–9126. [Online]. Available: https://doi.org/10.1109/ICASSP43922.2022.9746610

[13] G. Enzner and P. Vary, "Frequency-domain adaptive kalman filter for acoustic echo control in hands-free telephones," *Signal Process.*, vol. 86, no. 6, pp. 1140–1156, 2006. [Online]. Available: https://doi.org/10.1016/j.sigpro.2005.09.013

[14] E. A. Habets, "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, no. 2.4, p. 1, 2006.

[15] Y. Sun, L. Yang, H. Zhu, and J. Hao, "Funnel deep complex u-net for phase-aware speech enhancement," in *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, H. Hermansky, H. Cernocký, L. Burget, L. Lamel, O. Scharenborg, and P. Motlícek, Eds. ISCA, 2021, pp. 161–165. [Online]. Available: https://doi.org/10.21437/Interspeech.2021-10

[16] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, 2019. [Online]. Available: https://doi.org/10.1109/TASLP.2019.2915167

[17] Microsoft., "(2023) icassp acoustic echo cancellation challenge." [Online]. Available: https://www.microsoft.com/en-us/research/academic-program/acoustic-echo-cancellation-challenge-icassp-2023/

[18] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE transactions on acoustics, speech, and signal processing*, vol. 24, no. 4, pp. 320–327, 1976.

[19] J. Liu and X. Zhang, "ICCRN: inplace cepstral convolutional recurrent neural network for monaural speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*. IEEE, 2023, pp. 1–5. [Online]. Available: https://doi.org/10.1109/ICASSP49357.2023.10096918

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[21] 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 *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[22] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.