INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Residual Echo and Noise Cancellation with Feature Attention Module and Multi-domain Loss Function

Jianjun Gu[1,2], *Longbiao Cheng*[1,2], *Xingwei Sun*[1,2], *Junfeng Li*[1,2], *Yonghong Yan*[1,2,3]

[1]Institute of Acoustics, Chinese Academy of Sciences, China
[2]University of Chinese Academy of Sciences, China
[3]Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, China
`{gujianjun, chenglongbiao, sunxingwei, lijunfeng, yanyonghong}@hccl.ioa.ac.cn`

Abstract

For real-time acoustic echo cancellation in noisy environments, the classical linear adaptive filters (LAFs) can only remove the linear components of acoustic echo. To further attenuate the non-linear echo components and background noise, this paper proposes a deep learning-based residual echo and noise cancellation (RENC) model, where multiple inputs are utilized and weighted by a feature attention module. More specifically, input features extracted from the far-end reference and the echo estimated by the LAF are scaled with time-frequency attention weights, depending on their correlation with the residual interference in LAF's output. Moreover, a scale-independent mean square error and perceptual loss function are further suggested for training the RENC model. Experimental results validate the efficacy of the proposed feature attention module and multi-domain loss function, which achieve an 8.4%, 14.9% and 29.5% improvement in perceptual evaluation of speech quality (PESQ), scale-invariant signal-to-distortion ratio (SI-SDR) and echo return loss enhancement (ERLE), respectively.

Index Terms: Acoustic echo cancellation, residual echo suppression, deep learning

1. Introduction

In many telecommunication systems, acoustic echo is generated when the microphone picks up the loudspeaker's output then transfers a delayed version of the far-end signal back to the far end, which severely degrades the communication quality. Traditional acoustic echo cancellation (AEC) algorithms are based on the LAF, such as normalized least mean square (NLMS) updating rule [1] and Kalman filter [2], and seek for an estimation of the linear acoustic echo path from the loudspeaker to the microphone. Note that nonlinear distortions and background noise, which commonly exist in real-world scenarios, are not considered in these algorithms.

To address this problem, residual echo suppressors (RES) [3–8] and post-filters [9–11] based on traditional signal-processing-based methods have been proposed to further enhance the LAF's output. However, the performance of these methods degrades rapidly when the echo and noise become heavy. Inspired by the success of the deep neural networks (DNNs) in many fields, pure deep learning-based AEC methods with different network architectures have been investigated recently [12–16]. Benefiting from the powerful nonlinear system modeling ability of DNNs, these methods can simultaneously cancel all the interference signals and directly recover the near-end speech from the far-end reference and microphone signal. However, their generality may be problematic because too many acoustic environments (e.g., different room impulse responses, noise conditions, and equipment conditions) need to be modeled by the DNNs.

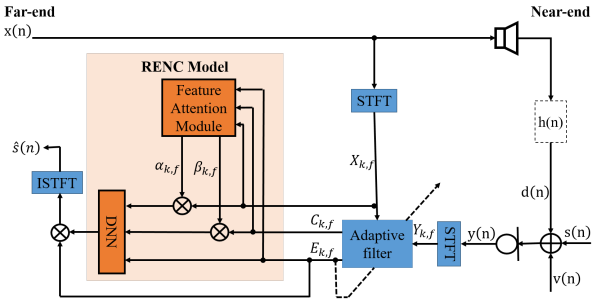


Figure 1: *Diagram of AEC system with proposed RENC model.*

Recently, DNNs have also been introduced into RES to remove the residual interference at the LAF's output [17–21], having better generalization ability thanks to the preprocessing of the LAF [19]. Inspired by the performance improvement achieved by utilizing multiple input features [18], these DNN based RES usually estimate the near-end speech from the combination of the LAF's output and other features (e.g., far-end reference and echo estimated by the LAF). It should be noted that these input features play a quite different role in the DNN. The LAF's output can provide information about the target near-end speech and all the residual interference needed to be removed. By comparison, the other inputs, i.e., the far-end signal and estimated echo, can only provide information concerning the residual echo in LAF's output, which will become useless for the DNN when dealing with the time-frequency (TF) bins where the LAF has canceled almost all the echo components with only background noise remaining. Therefore, the DNN needs to focus on different features depending on the situation. Unfortunately, all the input features are simply concatenated in the previous methods and contribute equally in all scenarios.

In this paper, we propose a deep learning-based RENC model with a feature attention module and multi-domain loss function. More specifically, the RENC model aims to estimate a gain function for recovering the near-end clean speech from the LAF's output signal. To make the model focus on different features depending on the context, a feature attention module is designed to generate the time-frequency weights for the input features extracted from the far-end signal and the estimated echo. Scale-independent mean square error (SI-MSE) is further proposed, and used for training RENC model together with the perceptual loss function suggested in [27], in which the SI-MSE loss overcomes the drawbacks of standard mean square

http://dx.doi.org/10.21437/Interspeech.2021-538

error (MSE) loss and the perceptual loss can further achieve a considerable improvement in speech perceptual quality. Experimental results in both single-talk and double-talk scenarios indicate that our proposed methods significantly outperform the comparative methods in terms of PESQ, SI-SDR, and ERLE.

2. Problem formulation

The general problem setting is illustrated in Fig.1, the near-end microphone signal $y(n)$ is a mixture of near-end speech $s(n)$, acoustic echo $d(n)$ and background noise $v(n)$:

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

where the acoustic echo $d(n)$ is generated by convolving the speaker output signal with a room impulse response (RIR) $h(n)$. Due to the limitations of components and the mechanical vibrations transmitted from the loudspeaker to the microphone, the speaker output signal is commonly regarded as a nonlinearly distorted version of the far-end reference signal $x(n)$. The objective of AEC system is to recover the near-end speech $s(n)$ with the priori knowledge of $y(n)$ and $x(n)$.

To remove the acoustic echo signal, the LAF based methods are commonly adopted to estimate the echo path adaptively with the signals $y(n)$ and $x(n)$ in time-frequency domain. After that, the estimated linear echo signal $C(k, f)$ can be calculated and canceled from $Y(k, f)$ to obtain the LAF's output $E(k, f)$:

$$E(k, f) = Y(k, f) - C(k, f), \tag{2}$$

where $Y(k, f)$ denotes the short time Fourier transform (STFT) spectrogram of $y(n)$, k and f indicate the frame and frequency bin index, respectively.

As the LAF removes only the linear echo components, the LAF's output $E(k, f)$ still contains residual echo and background noise. Usually, a post-processing module is required to further suppress the residual interference signals without distorting the near-end speech.

3. Proposed method

3.1. Overall structure

This paper proposes a deep learning-based RENC model to recover the near-end speech signal from the LAF's output. As demonstrated by the overall structure in Fig.2, the input features are firstly weighted by the feature attention module before being fed into the DNN, which is composed of two stacked gated recurrent unit (GRU) [23, 24] layers and a dense layer with sigmoid activation. Then, the DNN utilizes the weighted features to estimate the magnitude spectrum gain function $G(k, f)$ for the LAF's output to recover the estimated spectrum of near-end speech, which can be described as:

$$\hat{S}(k, f) = G(k, f) * E(k, f). \tag{3}$$

After that, the time-domain speech signal is recovered with inverse STFT process. GRUs are selected for their strong ability in modeling the sequences and high computational efficiency.

3.2. Feature attention module

The RENC model utilizes multiple input signals, including the LAF's output $E(k, f)$, the estimated linear echo $C(k, f)$ and the far-end reference signal $X(k, f)$, which has been proven to be an effective strategy [18]. The previous DNN-based RES methods normally concatenate these signals directly as input features, paying little attention to their difference. While $E(k, f)$ is required constantly to recover the near-end target, $X(k, f)$ and $C(k, f)$ are important for residual echo suppression but useless for noise cancellation. That is, the importance of $X(k, f)$ and $C(k, f)$ depends on their correlation with the residual interference. Therefore, we proposed a feature attention module to determine the weights of $X(k, f)$ and $C(k, f)$.

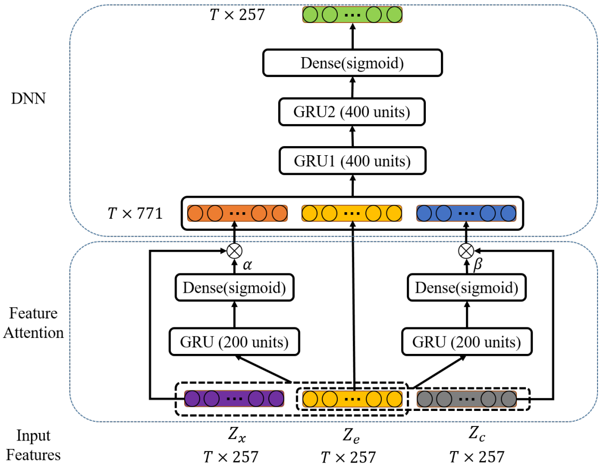


Figure 2: *Network architecture of the proposed RENC model.*

More specifically, the input features $Z_x(k, f)$, $Z_c(k, f)$ and $Z_e(k, f)$ are firstly calculated from the corresponding spectrum signals $X(k, f)$, $C(k, f)$ and $E(k, f)$ with log-power operators and normalized with online frequency-dependent normalization [25]. The feature attention module is based on a GRU layer followed by a dense layer with sigmoid activation as shown in Fig.2. The attention weights $\alpha(k, f)$ and $\beta(k, f)$ for $Z_x(k, f)$ and $Z_c(k, f)$ are estimated from their combination with $Z_e(k, f)$. Then, the attention weighted features $Z_x^{att}(k, f)$ and $Z_c^{att}(k, f)$ are calculated as:

$$\begin{aligned} Z_x^{att}(k, f) &= \alpha(k, f) * Z_x(k, f), \\ Z_c^{att}(k, f) &= \beta(k, f) * Z_c(k, f). \end{aligned} \tag{4}$$

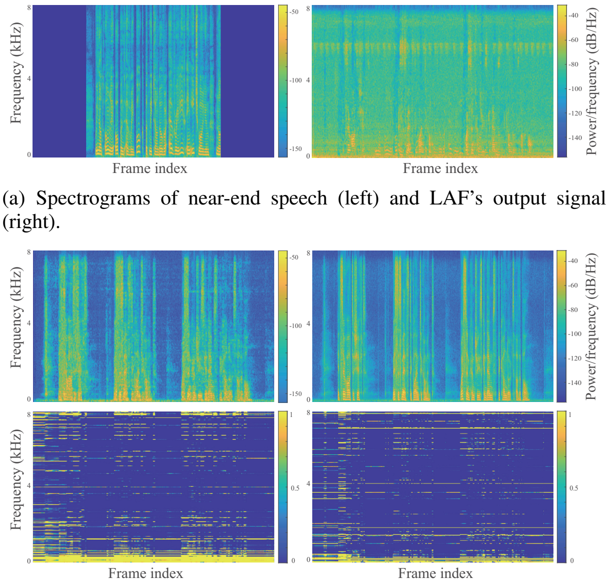
Finally, the input features for the DNN are obtained by concatenating $Z_x^{att}(k, f)$, $Z_e(k, f)$ and $Z_c^{att}(k, f)$ together along frequency dimension.

Fig.3 demonstrates the spectrograms and feature attention weights of an example processed by the proposed feature attention module. By comparing the LAF's output with the clean target in Fig.3 (a), it can be seen that residual echo mainly distributes in the low-frequency part while background noise exists in almost all the frequency bands. As we expected, the feature attention module applies significantly larger weights to the low-frequency part of features than the other frequency parts, as shown in Fig.3 (b). The non-speech periods of signals are also being scaled with relatively small weights since they cannot provide any valuable information.

3.3. Multi-domain loss function

In the model training process, MSE between target and estimated magnitude spectrum is widely used as loss function, which is calculated as:

$$L_{\text{MSE}}(S, \hat{S}) = E\{(|S(k, f)| - |\hat{S}(k, f)|)^2\}, \tag{5}$$



(a) Spectrograms of near-end speech (left) and LAF's output signal (right).

(b) Spectrograms (top) and attention weights (bottom) of estimated echo (left) and far-end signal (right).

Figure 3: *An example of feature attention weights.*

where $E\{\cdot\}$ denotes the mean of all the time-frequency components, and $|\cdot|$ means absolute operation. However, this MSE loss function has two drawbacks. Firstly, the MSE value may be affected by the energy level of the training samples. The training samples with large energy tend to be optimized with priority as the optimization of such samples brings a more obvious loss decrease. On the contrary, the samples with smaller energy cannot be optimized well as it brings a minor decrease of the final batch-based loss. Secondly, it may be affected by the different energy levels of the estimated and target spectrum, although the perceptual results are no different after global energy scaling. Therefore, we proposed a SI-MSE loss function on magnitude spectrum, which is calculated as:

$$L_{\text{SI-MSE}}(S,\hat{S}) = \frac{E\{(|\tilde{S}(k,f)| - |\hat{S}(k,f)|)^2\}}{E\{|\tilde{S}(k,f)|^2\}}, \qquad (6)$$

where

$$\tilde{S}(k,f) = \frac{E\{|S(k,f)| * |\hat{S}(k,f)|\}}{E\{|S(k,f)|^2\}} * S(k,f). \qquad (7)$$

Specifically, the normalized target spectrum $\tilde{S}(k,f)$ is obtained by projecting the target spectrum onto the estimated spectrum as described in Eq.7. Then, the SI-MSE loss is calculated in Eq.6 by normalizing the MSE between the magnitude of $\tilde{S}(k,f)$ and $\hat{S}(k,f)$ with the magnitude energy of $\tilde{S}(k,f)$.

To further improve the speech perceptual quality of the estimated speech, we also adopted a perceptual loss function for training the RENC model. The perceptual loss function is adapted from the simplified PESQ algorithm with both gain and frequency equalization in [27] which is suitable for gradient-based training. Finally, the hybrid multi-domain loss function is obtained:

$$L(S,\hat{S}) = \lambda L_{\text{SI-MSE}}(S,\hat{S}) + L_{\text{PESQ}}(S,\hat{S}), \qquad (8)$$

where $L_{\text{PESQ}}(S,\hat{S})$ is the perceptual loss function and the parameter λ is used to balance the values of the loss function.

4. Experimental results

4.1. Data preparation

In our experiments, we utilize the AEC Challenge dataset [28] which is composed of both synthetic data and real recorded data. There are 10,000 synthetic utterance samples representing single-talk, double-talk, near-end noise, far-end noise, and various nonlinear distortion situations, which can be directly used for model training. The real recordings are from more than 2,500 different real environments, audio devices, and human speakers in both single-talk and double-talk scenarios. As the real recorded data has no clean near-end target, only the far-end single-talk records are used to generate double-talk mixtures by mixing them with the near-end speech from the synthetic dataset. Almost 3,000 far-end single-talk records with or without echo path change are selected from the real recorded dataset based on their recording quality, in which 2800, 100, and 100 records are used to generate double-talk mixtures for training, validation, and testing set, separately. Each far-end single-talk record in the training and validation set is mixed with ten different near-end signals at the signal to echo and noise ratio (SENR) level randomly chosen from -25dB to 25dB after being cut or zero-extended to the same length of the near-end signal. Similarly, the selected 100 testing records are used to generate 400 mixtures at five different SENR levels $\{-15,-10,-5,0,5\}$ dB randomly for testing in double-talk situations, while the original 100 records without mixing near-end speech are used for testing in far-end single-talk situations. The training utterance samples generated from the real recorded dataset are used for model training with the utterance samples in the synthetic dataset. The SENR here is defined as:

$$\text{SENR} = 10\log_{10}\frac{E\{s(n)^2\}}{E\{[d(n)+v(n)]^2\}}. \qquad (9)$$

4.2. Comparison Method and Training Setup

To demonstrate the effectiveness of each component proposed in this paper, nine systems are adopted for comparison in total. Kalman denotes the system with only the Kalman filter proposed in [22]. BLSTM denotes the pure DNN based AEC algorithm in [12], which trained a recurrent neural network with bidirectional long short-term memory (BLSTM) to estimate the ideal ratio mask [29]. The rest of the systems are cascaded methods combining Kalman filter and DNN-based post-filter and are denoted as Kal. + •. Among them, FCN denotes the fully connected network (FCN) model in [18] with phase-sensitive filter (PSF) [30] output. FC-MSE and FC-SIMSE denote systems without feature attention and trained with conventional MSE loss and the proposed SI-MSE loss, respectively. FA-SIMSE denotes the proposed system trained with SI-MSE loss only. The FA-SIMSE is further trained with the loss in Eq.8 to evaluate the effect of the perceptual loss, and this method is denoted as $+\mathcal{L}_{\text{PESQ}}$. Further, to determine the features to be weighted in the feature attention module, another two algorithms denoted as F-FA-SIMSE and E-FA-SIMSE are adopted for comparison, which only apply attention weights to the far-end signal feature or the estimated echo feature, respectively.

In the model training process, all the input signals sampled at 16 kHz are divided into 32-ms frames with an 8-ms frame shift. For each frame, a 512 point STFT is performed to generate the speech spectrum with 257 dimensions. All the models are trained for 50 epochs using Adam optimizer with an initial learning rate of 0.001 and decay 90% when the validation loss increases. The λ in Eq.8 is set to 50 in our experiments.

Table 1: *PESQ and SI-SDR values on double-talk records*

Method		SENR(dB)(Double-Talk)									
		-15		-10		-5		0		5	
		PESQ	SISDR	PESQ	SISDR	PESQ	SISDR	PESQ	SISDR	PESQ	SISDR
Noisy		1.188	-15.079	1.492	-10.051	1.688	-4.985	2.046	0.001	2.345	5.000
Kalman		1.567	-8.717	1.895	-3.449	2.192	1.574	2.520	5.967	2.834	10.044
BLSTM [12]		1.438	-1.891	1.839	2.561	2.075	6.234	2.339	9.609	2.690	13.645
Kal. +	FCN [18]	1.574	-2.893	1.918	2.195	2.294	6.494	2.667	9.604	2.951	11.678
	FC-MSE	1.494	0.740	1.858	4.675	2.206	8.409	2.555	10.968	2.845	13.476
	FC-SIMSE	1.548	1.815	1.998	5.847	2.322	9.200	2.660	11.473	2.925	13.781
	FA-SIMSE	**1.653**	2.310	2.074	**6.120**	2.384	**9.604**	2.725	11.788	2.979	14.050
	$+\mathcal{L}_{\mathrm{PESQ}}$	1.640	**2.405**	**2.080**	6.108	**2.402**	9.597	**2.760**	**11.797**	**3.003**	**14.055**

Table 2: *ERLE values on both far-end single-talk records and single-talk periods of double-talk records*

Methods		SENR(dB)(Double-Talk)					Single-Talk	Average
		-15	-10	-5	0	5		
Kalman		7.754	7.844	7.966	7.900	8.471	8.523	8.076
BLSTM [12]		21.656	22.018	22.467	22.534	24.181	23.483	22.723
Kal. +	FCN [18]	21.041	21.374	21.741	21.907	23.480	23.273	22.136
	FC-MSE	34.180	36.541	36.675	36.697	37.128	23.936	34.193
	FC-SIMSE	38.558	40.249	39.411	40.490	40.871	27.399	37.830
	FA-SIMSE	40.685	43.561	42.944	43.109	42.851	30.051	40.534
	$+\mathcal{L}_{\mathrm{PESQ}}$	**45.800**	**47.650**	**47.379**	**46.980**	**45.454**	**32.335**	**44.266**

Table 3: *Evaluation results of weighting different features*

Methods (Kal. +)	SENR(dB)(Double-Talk)									Single-Talk
	-15			-5			5			
	PESQ	SISDR	ERLE	PESQ	SISDR	ERLE	PESQ	SISDR	ERLE	ERLE
F-FA-SIMSE	1.610	2.059	**40.739**	2.360	9.298	41.878	2.950	13.965	42.007	27.795
E-FA-SIMSE	1.571	1.678	38.782	2.302	9.097	40.284	2.930	13.772	39.359	27.361
FA-SIMSE	**1.653**	**2.310**	40.685	**2.384**	**9.604**	**42.944**	**2.979**	**14.050**	**42.851**	**30.051**

4.3. Performance Evaluation and Analysis

The performance of all systems is evaluated in terms of ERLE [31] for both the simulated far-end single-talk records and single-talk periods of the simulated double-talk records, PESQ [26] SI-SDR [32] for the simulated double-talk records.

Table I and Table II show the evaluation results. In general, the proposed algorithm using feature attention and multi-domain loss achieves the best performance on the overall three metrics. BLSTM performs worse than all the proposed cascaded systems, although it is non-causal, illustrating the superiority of combining LAF and DNN based post-filter. All the variants of the proposed method outperform FCN in terms of SISDR and ERLE since GRU can utilize relations between time series much better than fully connected network. Compared with FC-MSE, FC-SIMSE has a better performance on PESQ, SISDR, and ERLE. Therefore, the effectiveness of SI-MSE can be confirmed. FA-SIMSE has significantly better performance than FC-SIMSE in all conditions, which is benefitting from the feature attention proposed in this work. Results of the FA-SIMSE and FA-SIMSE+$\mathcal{L}_{\mathrm{PESQ}}$ manifest that the $\mathcal{L}_{\mathrm{PESQ}}$ would provide a further improvement on not only the PESQ scores but also the SI-SNR and ERLE scores, which should lead to a better perceptual quality for the recovered near-end speech.

Table III shows the evaluation results of proposed algorithms with different weighting methods for the input features. FA-SIMSE performs best under almost all the conditions in terms of PESQ, SI-SDR and ERLE, validating the necessity of weighting both far-end signal and estimated echo features. Note that F-FA-SIMSE outperforms E-FA-SIMSE in all the metrics. A possible reason for this is that the far-end signal contains more irrelevant information that needs to be blocked than the estimated echo.

5. Conclusion

This paper proposes a deep learning-based RENC model with a feature attention module and multi-domain loss function. The feature attention mechanism is proposed to obtain better performance by weighting the input features adaptively depending on the scenarios. Multi-domain loss function consisting of the SI-MSE loss and perceptual loss is further proposed for training, which could overcome the drawbacks of traditional MSE loss and further improve the speech perceptual quality. Evaluation results confirm the effectiveness of each proposed component in this work. As future work, we intend to apply the proposed future attention module to the network structures that are more appropriate for removing the residual echo and noise.

6. Acknowledgment

This research is partially supported by the National Key Research and Development Program (No.2020YFC2004100) and the National Natural Science Foundation of China (No.11911540067).

7. References

[1] J. Benesty and S. L. Gay, "An improved pnlms algorithm," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, 2002, pp. II–1881–II–1884.

[2] C. Paleologu, J. Benesty, and S. Ciochină, "Study of the general kalman filter for echo cancellation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 8, pp. 1539–1549, 2013.

[3] E. A. P. Habets, S. Gannot, I. Cohen, and P. C. W. Sommen, "Joint dereverberation and residual echo suppression of speech signals in noisy environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1433–1451, 2008.

[4] M. L. Valero, E. Mabande, and E. A. P. Habets, "Signal-based late residual echo spectral variance estimation," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 5914–5918.

[5] N. K. Desiraju, S. Doclo, M. Buck, and T. Wolff, "Online estimation of reverberation parameters for late residual echo suppression," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 77–91, 2020.

[6] S. Gustafsson, R. Martin, P. Jax, and P. Vary, "A psychoacoustic approach to combined acoustic echo cancellation and noise reduction," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 245–256, 2002.

[7] S. Gustafsson, R. Martin, and P. Vary, "Combined acoustic echo control and noise reduction for hands-free telephony," *Signal processing*, vol. 64, no. 1, pp. 21–32, 1998.

[8] A. S. Chhetri, M. Ananda, J. Stokes, and J. C. Platt, "Regression-based residual acoustic echo suppression," in *International Workshop on Acoustic Echo and Noise Control IWAENC '05, Eindhoven, Netherlands*, September 2005.

[9] K. Nathwani, "Joint acoustic echo and noise cancellation using spectral domain kalman filtering in double-talk scenario," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018, pp. 1–330.

[10] V. Turbin, A. Gilloire, and P. Scalart, "Comparison of three post-filtering algorithms for residual acoustic echo reduction," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 1997, pp. 307–310.

[11] F. Ykhlef and H. Ykhlef, "A post-filter for acoustic echo cancellation in frequency domain," in *2014 Second World Conference on Complex Systems (WCCS)*, 2014, pp. 446–450.

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

[13] H. Zhang, K. Tan, and D. Wang, "Deep Learning for Joint Acoustic Echo and Noise Cancellation with Nonlinear Distortions," in *Proc. Interspeech 2019*, 2019, pp. 4255–4259.

[14] A. Fazel, M. El-Khamy, and J. Lee, "Deep Multitask Acoustic Echo Cancellation," in *Proc. Interspeech 2019*, 2019, pp. 4250–4254.

[15] J.-H. Kim and J.-H. Chang, "Attention Wave-U-Net for Acoustic Echo Cancellation," in *Proc. Interspeech 2020*, 2020, pp. 3969–3973.

[16] Y. Zhang, C. Deng, S. Ma, Y. Sha, H. Song, and X. Li, "Generative Adversarial Network Based Acoustic Echo Cancellation," in *Proc. Interspeech 2020*, 2020, pp. 3945–3949.

[17] A. Fazel, M. El-Khamy, and J. Lee, "Cad-aec: Context-aware deep acoustic echo cancellation," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6919–6923.

[18] G. Carbajal, R. Serizel, E. Vincent, and E. Humbert, "Multiple-input neural network-based residual echo suppression," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 231–235.

[19] C. Zhang and X. Zhang, "A Robust and Cascaded Acoustic Echo Cancellation Based on Deep Learning," in *Proc. Interspeech 2020*, 2020, pp. 3940–3944.

[20] L. Pfeifenberger and F. Pernkopf, "Nonlinear Residual Echo Suppression Using a Recurrent Neural Network," in *Proc. Interspeech 2020*, 2020, pp. 3950–3954.

[21] H. Chen, T. Xiang, K. Chen, and J. Lu, "Nonlinear Residual Echo Suppression Based on Multi-Stream Conv-TasNet," in *Proc. Interspeech 2020*, 2020, pp. 3959–3963.

[22] S. Braun and E. A. P. Habets, "Online dereverberation for dynamic scenarios using a kalman filter with an autoregressive model," *IEEE Signal Processing Letters*, vol. 23, no. 12, pp. 1741–1745, 2016.

[23] K. Cho, B. van Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," in *Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, 2014.

[24] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014.

[25] Y. Xia, S. Braun, C. K. A. Reddy, H. Dubey, R. Cutler, and I. Tashev, "Weighted speech distortion losses for neural-network-based real-time speech enhancement," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 871–875.

[26] 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, 2001, pp. 749–752.

[27] J. M. Martin-Doñas, A. M. Gomez, J. A. Gonzalez, and A. M. Peinado, "A deep learning loss function based on the perceptual evaluation of the speech quality," *IEEE Signal Processing Letters*, vol. 25, no. 11, pp. 1680–1684, 2018.

[28] K. Sridhar, R. Cutler, A. Saabas, T. Parnamaa, H. Gamper, S. Braun, R. Aichner, and S. Srinivasan, "Icassp 2021 acoustic echo cancellation challenge: Datasets and testing framework," *arXiv preprint arXiv:2009.04972*, 2020.

[29] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[30] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 708–712.

[31] 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.

[32] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr – half-baked or well done?" in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 626–630.