NEURAL CASCADE ARCHITECTURE FOR JOINT ACOUSTIC ECHO AND NOISE SUPPRESSION

Hao Zhang[1], *DeLiang Wang*[1,2]

[1]Department of Computer Science and Engineering, The Ohio State University, USA
[2]Center for Cognitive and Brain Sciences, The Ohio State University, USA

ABSTRACT

In this paper, we propose a neural cascade architecture for joint acoustic echo and noise suppression. The proposed cascade architecture consists of two modules. A convolutional recurrent network (CRN) is employed in the first module for complex spectral mapping. The output is then 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. Evaluation results show that the proposed method effectively suppresses acoustic echo and noise while preserving good speech quality, and significantly outperforms related methods.

Index Terms— Neural cascade architecture, acoustic echo suppression, deep learning, complex spectral mapping

1. INTRODUCTION

Acoustic echo cancellation (AEC) plays an essential role in smart speaker and speech communication systems and has received significant attention for several decades [1, 2]. Conventionally, acoustic echo is removed by identifying an acoustic impulse response between loudspeaker and microphone using adaptive algorithms such as normalized least mean square (NLMS) [3]. However, the presence of double-talk and background noise, especially non-stationary noise, disrupts the convergence of these algorithms. Moreover, most of the traditional AEC algorithms are fundamentally linear systems, which cannot deal with nonlinear distortions introduced to the recorded echo due to the poor quality of devices such as amplifiers and loudspeakers [4, 5].

Deep learning has been utilized recently for addressing AEC problems due to its capacity in modeling nonlinear relations. Lee et al. [6] use a neural network as a residual echo suppressor to remove the nonlinear components of echo. Zhang and Wang [7] formulate AEC as a supervised speech separation problem and achieve echo suppression by extracting the near-end speech from a microphone recording. Carbajal et al. [8] achieve echo removal by utilizing a multi-input neural network based method to estimate phase-sensitive masks. Early studies focus on magnitude enhancement and use magnitude masks as the training target [6, 8, 7]. Later, complex-domain estimation is employed for phase enhancement to improve the quality of estimated near-end speech. Convolutional recurrent neural network (CRN) based methods that use complex spectral mapping for echo suppression are introduced in [9, 10]. AEC methods that use complex-valued neural networks for phase-aware enhancement are described in [11, 12].

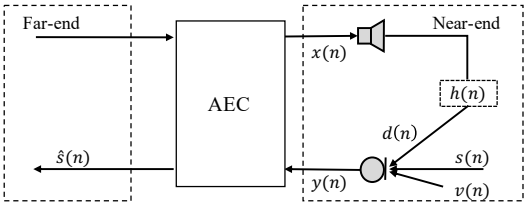

Fig. 1. A typical acoustic echo cancellation system.

Recently, AEC Challenges [13, 14] show that various deep learning architectures can be utilized to address AEC problems. There is a trend of combining traditional and deep learning based 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 [15, 16, 11]. Moreover, multiple neural networks have been combined to perform joint echo and noise suppression [17, 18, 19].

Although deep learning based methods have been successfully applied to AEC with convincing results, there exists a tradeoff between echo suppression and near-end speech quality. Franzen et al. reveal this tradeoff in [20] and provide insights to the choice of inputs for better echo removal [21]. A suppression loss that balances this tradeoff is introduced in [22]. Ivry et al. propose a system with a design parameter that allows a tunable tradeoff between echo suppression and near-end speech distortion [23].

Inspired by a recent neural cascade architecture for speech enhancement [24], we propose a neural cascade architecture for joint 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 multiple sequential training processes 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 for generating time-domain near-end speech. Hence the cascade architecture leverages the advantages of the two modules and obtains robust magnitude estimation as well as phase enhancement.

The remainder of this paper is organized as follows. Section 2 describes the problem and introduces magnitude and complex-

domain estimation. Section 3 presents the proposed neural cascade architecture. Evaluation and comparison results are given in Section 4. Section 5 concludes this paper.

2. PROBLEM FORMULATION

2.1. Signal model

As is shown in Fig. 1, 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, the echo signal is generated by convolving a loudspeaker signal (or far-end signal $x(n)$ when there is no nonlinear distortion) with a room impulse response (RIR) between loudspeaker and microphone $h(n)$.

Joint acoustic echo and noise suppression aims to cancel echo and noise and send only the near-end speech to the far end. Deep learning based methods formulate the problem as a supervised speech separation problem and work by directly extracting the near-end speech from the microphone recording.

2.2. Magnitude mask estimation vs. complex-domain estimation

The echo and noise suppression performance depends on the accuracy of the estimated magnitude while the quality of extracted near-end speech is highly related to the phase information. The magnitude mask estimation based methods utilize ratio masks as the training targets. The value range of these masks is bounded between $[0, 1]$, which is easier to predict and usually leads to better echo and noise suppression performance. The complex-domain methods emphasize the importance of phase estimation and jointly estimate magnitude and phase, resulting in improvement in speech quality. However, the value range of the real and imaginary targets used in the complex-domain estimation methods, either complex spectral mapping or complex ratio mask, are unbounded. Although a few techniques have been proposed to bound the output range, it is still harder to achieve a robust magnitude estimate compared to the magnitude mask estimation based methods.

3. NEURAL CASCADE ARCHITECTURE

The proposed neural cascade architecture consists of a complex module and a magnitude mask module, as is 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 a robust magnitude estimate. We have also explored other cascading mechanisms such as using magnitude mask estimation before complex-domain estimation. The architecture presented in this paper achieves the best overall performance.

3.1. 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 within a T-F unit at time t and frequency f, respectively, subscript $_r$ and $_i$ denote the real and imaginary spectrograms of the corresponding signals. The enhanced magnitude and phase are then calculated, respectively, as:

$$\hat{S}'_m(t,f) = \sqrt{\hat{S}'^2_r(t,f) + \hat{S}'^2_i(t,f)} \tag{2}$$

$$\theta_{\hat{S}'}(t,f) = \arctan\left(\hat{S}'_i(t,f)/\hat{S}'_r(t,f)\right) \tag{3}$$

The CRN is 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 [25] except that our CRN has four input channels.

3.2. Magnitude mask module

The estimated $\hat{S}'_m(t,f)$, together with $Y_m(t,f)$ and $X_m(t,f)$ are fed to the magnitude mask module to predict a T-F mask $M(t,f)$ using a long short-term memory network (LSTM). The estimated magnitude spectrogram is obtained from:

$$\hat{S}_m(t,f) = M(t,f) \odot Y_m(t,f) \tag{4}$$

where $\odot$ denotes element-wise multiplication, subscript $_m$ denotes the magnitude spectrogram of the corresponding signals.

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

$$\hat{s}(n) = \text{iSTFT}\left(\hat{S}_m(t,f), \theta_{\hat{S}'}(t,f)\right) \tag{5}$$

The LSTM has four hidden layers with 300 units in each layer. The output layer is a fully connected layer. Since the value range of the output is $[0, 1]$, the sigmoid function is used as the activation function in the output layer.

3.3. Loss functions

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

Following [26], we define the first loss L_{complex} as the real, imaginary, and magnitude difference between $\hat{S}'(t,f)$ and $S(t,f)$:

$$\begin{aligned} L_{\text{complex}} = &\frac{1}{TF}\sum_{t,f}(|\hat{S}'_r(t,f) - S_r(t,f)|^2 \\ &+ |\hat{S}'_i(t,f) - S_i(t,f)|^2 + |\hat{S}'_m(t,f) - S_m(t,f)|^2) \end{aligned} \tag{6}$$

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

$$L_{\text{mag-mask}} = \frac{1}{TF}\sum_{t,f}\left(M(t,f) \odot Y_m(t,f) - S_m(t,f)\right)^2 \tag{7}$$

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 a single loss function:

$$L_{\text{combined}} = \lambda L_{\text{complex}} + (1-\lambda) L_{\text{mag-mask}} \tag{8}$$

where λ is a coefficient for combining the two losses. We empirically select $\lambda = \frac{2}{3}$ based on the performance on the validation data.

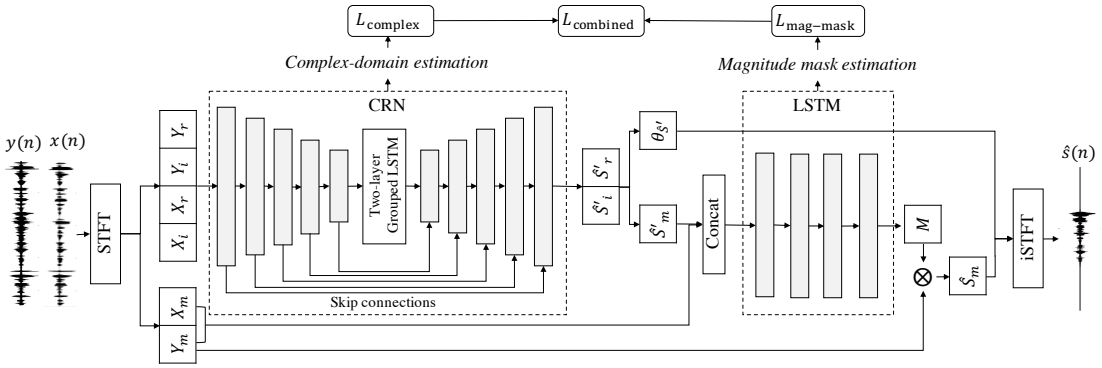

Fig. 2. Diagram of the proposed neural cascade architecture for joint echo and noise suppression. The first module employs a CRN for complex spectral mapping, the output is concatenated with original inputs and fed to an LSTM to predict T-F masks. Subscripts $_r$, $_i$, and $_m$ denote real, imaginary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

4. EXPERIMENTAL RESULTS

4.1. Experiment settings

We follow the experimental setup introduced in [9, 15] and perform experiments in situations with double-talk, background noise, and nonlinear distortions. We randomly choose 100 pairs of speakers from the 630 speakers in the TIMIT dataset [27] as the near-end and far-end speakers, respectively. Three randomly selected utterances from a far-end speaker is concatenated to form a far-end signal. A randomly chosen near-end speech is extended to the same length by adding zeros at both the front and rear of the signal. To achieve a noise-independent model, we use 10000 noises from a sound effect library (available at http://www.sound-ideas.com) for creating training mixtures (see [28]). Operational room noise (oproom) speech shaped noise (SSN) from NOISEX-92 dataset [29], babble noise from the Auditec CD (http://www.auditec.com), and white noise are used for creating test mixtures. Note that the noises used for testing are different from those for training.

RIRs are simulated using the image method [30]. To investigate RIR generalization, we simulate 20 rooms of different sizes $a \times b \times c$ m^3, where $a \in \{4, 6, 8, 10\}, b \in \{5, 7, 9, 11, 13\}$, and $c = 3$ for the training mixtures. Ten pairs of random positions in each room are simulated to generate RIRs for the loudspeaker and near-end speaker. The reverberation time (T_{60}) is randomly selected from $\{0.2, 0.3, 0.4\}$ s. Therefore, in total, 200 pairs of RIRs are synthesized to create training mixtures. Two rooms of untrained sizes, $3 \times 4 \times 3$ m and $11 \times 14 \times 3$ m are simulated to generating RIRs for testing. Ten pairs of RIRs are generated in each of the rooms, which are denoted as RIR1s and RIR2s, respectively. The nonlinear distortions introduced by a power amplifier and a loudspeaker are simulated by following the steps introduced in [6, 31].

We create 20000 training and 300 test mixtures. Each training mixture is created by first adding nonlinear distortions to a randomly chosen far-end signal and convolving it with a randomly chosen RIR 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 echo at a signal-to-echo ratio (SER) randomly chosen from $\{-6, -3, 0, 3, 6\}$ dB. Finally, a random cut from the 10000 noises is added to the mixture at a signal-to-noise ratio (SNR) randomly chosen from $\{8, 10, 12, 14\}$ dB. The SER and SNR are evaluated during double-talk periods. Test mixtures are created similarly but using different utterances, noises, RIRs, SERs, and SNRs.

Performance of the proposed method is evaluated in terms of ERLE for single-talk periods and perceptual evaluation of speech quality (PESQ) [32] for double-talk periods. ERLE is defined as:

$$\text{ERLE} = 10 \log 10 \left[\sum_n y(n)^2 / \sum_n \hat{s}^2(n) \right] \tag{9}$$

Evaluation results are presented as mean $\pm$ std.

4.2. Evaluation results

We first evaluate the proposed method and compare it with four deep learning based baseline methods. The evaluation results are provided in Table 1. Architectures of the LSTM method and the CRN method [9] are the same as the two modules used in our proposed architecture. The LSTM baseline is adopted from [7] by replacing the bidirectional LSTM in the original model with unidirectional LSTM. The multi-input residual echo suppression (MI-RES) method is a two-stage system that combines adaptive algorithm with neural network [8]. The LFM-NFM [15] is a cascaded AEC method that consists of two neural networks, which serve as a linear-filtering model (LFM) and a nonlinear-filtering model (NLM).

It is seen from the table that the proposed method consistently outperforms the other multi-stage methods. And compared with the LSTM based magnitude mask estimation and the CRN based complex spectral mapping, the proposed cascade architecture achieves better echo removal and speech quality. Results provided in Fig. 3 show that the performance of the proposed method generalizes well to untrained noises. Spectrograms of a test sample are given in Fig. 4. It is seen that the output of the proposed method approximates the target near-end speech and has less residual echo and noise compared to other methods.

4.3. Performance using different training strategies

We further evaluate the performance of the proposed cascade architecture trained using different strategies. The comparison results under RIR1s, 3.5 dB SER, 10 dB SNR, and white noise are given in Table 2. There are two reasonable masking strategies for the magnitude mask module in the proposed architecture, which are applying the estimated magnitude mask upon microphone signal Y_m or the estimated near-end speech $\hat{S}'_m$. And the models trained using these two strategies achieve comparable performance while the proposed method is slightly better than the other one. This is because the estimated $\hat{S}'_m$ has distortions in it, estimating $\hat{S}_m$ by applying a mask

Table 1. Performance in the presence of double-talk, white noise, nonlinear distortions and untrained RIRs (RIR1) with 10 dB SNR and different SER.

	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 [8]	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 [7]	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 [9]	35.64 ± 2.44	2.64 ± 0.18	36.87 ± 2.66	2.51 ± 0.18	37.50 ± 2.54	2.33 ± 0.21
LFM-NFM [15]	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	$\mathbf{53.43 \pm 6.43}$	$\mathbf{2.68 \pm 0.16}$	$\mathbf{53.75 \pm 6.31}$	$\mathbf{2.54 \pm 0.16}$	$\mathbf{53.00 \pm 7.56}$	$\mathbf{2.37 \pm 0.19}$

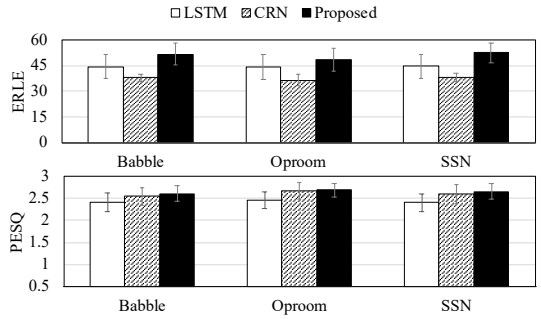

Fig. 3. ERLE and PESQ values in the presence of different untrained noises, SER = 3.5 dB, SNR = 10 dB.

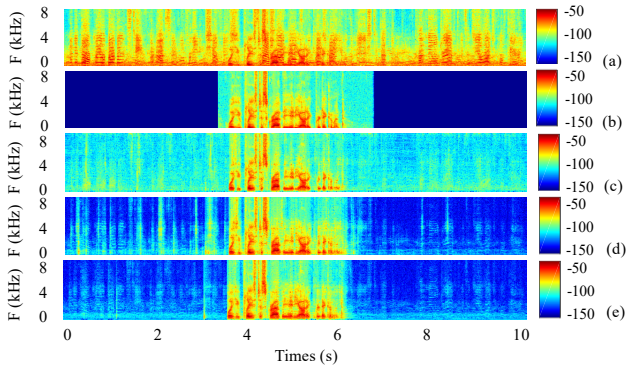

Fig. 4. Spectrograms of a test sample with 3.5 dB SER and babble noise at 10 dB SNR: (a) microphone signal, (b) target near-end speech, and outputs of (c) CRN, (d) LSTM, (e) Proposed.

upon $\hat{S}'_m$ could further distort the speech components. The fourth row of the table shows the results of the cascade architecture trained sequentially using separate loss functions. And the last row is the results of the proposed architecture trained by only optimizing the loss function at the final output, $L_{\text{mag-mask}}$. Comparison results illustrate that the strong performance of the proposed method is not only due to the neural network structure, but also benefits from the combined loss function and the training strategy.

4.4. Robustness test

The proposed method is further tested in situations with untrained speakers, untrained RIRs (RIR2), 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 RIR1s. The test mixtures under untrained larger rooms are generated using RIR2s. The echo path change is simulated by randomly select two pairs of RIRs from RIR1s 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 datasets. The results given in Table 3 indicate the strong robustness of the proposed method.

Table 2. Performance using different training strategies under 3.5 dB SER, 10 dB SNR and white noise.

3.5 dB SER	ERLE	PESQ
Unprocessed	0	1.96 ± 0.19
Proposed	$\mathbf{53.43 \pm 6.43}$	$\mathbf{2.68 \pm 0.16}$
Applying mask upon $\hat{S}'_m$	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 3. Performance under untrained speakers, RIRs, and echo path change conditions with 3.5 dB SER,10 dB SNR, white noise.

	ERLE	PESQ	
	Proposed	Unprocessed	Proposed
Untrained speaker	53.57 ± 6.30	1.95 ± 0.23	2.69 ± 0.19
Untrained larger room	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

5. CONCLUSIONS

We have proposed a neural cascade architecture for joint acoustic echo and noise suppression. The main idea is to leverage the advantages of complex spectral mapping and magnitude mask estimation to achieve joint phase and magnitude enhancement. 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 show that the proposed method outperforms other baseline methods and generalizes well to untrained scenarios.

6. ACKNOWLEDGEMENTS

This research was supported in part by an NIDCD grant (R01 DC012048) and the Ohio Supercomputer Center.

7. REFERENCES

[1] M. 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*, vol. 4, pp. 807–877. Elsevier, 2014.

[4] A. N. Birkett and R. A. Goubran, "Limitations of handsfree acoustic echo cancellers due to nonlinear loudspeaker distortion and enclosure vibration effects," in *Proceedings of WASPAA*. IEEE, 1995, pp. 103–106.

[5] M. I. Mossi, N. W. Evans, and C. Beaugeant, "An assessment of linear adaptive filter performance with nonlinear distortions," in *2010 ICASSP*. IEEE, 2010, pp. 313–316.

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

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

[8] G. Carbajal, R. Serizel, E. Vincent, and E. Humbert, "Multiple-input neural network-based residual echo suppression," in *Proceedings of INTERSPEECH*, 2018, pp. 231–235.

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

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

[11] 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 *ICASSP*. IEEE, 2021, pp. 121–125.

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

[13] K. Sridhar, R. Cutler, A. Saabas, T. Parnamaa, M. Loide, H. Gamper, S. Braun, R. Aichner, and S. Srinivasan, "ICASSP 2021 acoustic echo cancellation challenge: Datasets, testing framework, and results," in *ICASSP*. IEEE, 2021, pp. 151–155.

[14] R. Cutler, A. Saabas, T. Parnamaa, M. Loide, S. Sootla, M. Purin, H. Gamper, S. Braun, K. Sorensen, R. Aichner, et al., "Interspeech 2021 acoustic echo cancellation challenge," in *Proceedings of INTERSPEECH*, 2021.

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

[16] 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 *ICASSP*. IEEE, 2021, pp. 7133–7137.

[17] X. Shu, Y. Zhu, Y. Chen, L. Chen, H. Liu, C. Huang, and Y. Wang, "Joint echo cancellation and noise suppression based on cascaded magnitude and complex mask estimation," *arXiv preprint arXiv:2107.09298*, 2021.

[18] N. L. Westhausen and B. T. Meyer, "Acoustic echo cancellation with the dual-signal transformation LSTM network," in *ICASSP*. IEEE, 2021, pp. 7138–7142.

[19] 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 *ICASSP*. IEEE, 2021, pp. 146–150.

[20] J. Franzen, E. Seidel, and T. Fingscheidt, "AEC in a Netshell: on target and topology choices for FCRN acoustic echo cancellation," in *ICASSP*. IEEE, 2021, pp. 156–160.

[21] J. Franzen and T. Fingscheidt, "Deep residual echo suppression and noise reduction: A multi-input FCRN approach in a hybrid speech enhancement system," *arXiv preprint arXiv:2108.03051*, 2021.

[22] X. Zhou and Y. Leng, "Residual acoustic echo suppression based on efficient multi-task convolutional neural network," *arXiv preprint arXiv:2009.13931*, 2020.

[23] A. Ivry, I. Cohen, and B. Berdugo, "Deep residual echo suppression with a tunable tradeoff between signal distortion and echo suppression," in *ICASSP*. IEEE, 2021, pp. 126–130.

[24] H. Wang and D. L. Wang, "Neural cascade architecture with triple-domain loss for speech enhancement," *under journal review*, 2021.

[25] K. Tan and D. L. Wang, "Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement," in *ICASSP*. IEEE, 2019, pp. 6865–6869.

[26] Z. Q Wang, P. Wang, and D. L. Wang, "Complex spectral mapping for single-and multi-channel speech enhancement and robust asr," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 28, pp. 1778–1787, 2020.

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

[28] J. Chen, Y. Wang, S. E Yoho, D. L. Wang, and E. W. Healy, "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," *The Journal of the Acoustical Society of America*, vol. 139, no. 5, pp. 2604–2612, 2016.

[29] 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 communication*, vol. 12, no. 3, pp. 247–251, 1993.

[30] J. B. Allen and D. A. 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.

[31] D. Comminiello, M. Scarpiniti, L. A. Azpicueta-Ruiz, J. Arenas-García, and A. Uncini, "Functional link adaptive filters for nonlinear acoustic echo cancellation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1502–1512, 2013.

[32] 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 *ICASSP*. IEEE, 2001, vol. 2, pp. 749–752.