INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

A Two-stage Progressive Neural Network for Acoustic Echo Cancellation

Zhuangqi Chen[1,2,*], *Xianjun Xia*[1], *Cheng Chen*[1], *Xianke Wang*[1,3], *Yanhong Leng*[1], *Li Chen*[1], *Roberto Togneri*[4], *Yijian Xiao*[1], *Piao Ding*[1], *Shenyi Song*[1], *Pingjian Zhang*[2]

[1]RTC Lab, ByteDance, China
[2]SSE, South China University of Technology, Guangzhou, China
[3]EIC, Huazhong University of Science and Technology, Wuhan, China
[4]EECE, University of Western Australia, Perth, Australia

zqchen12cc@163.com, xxjpjj@mail.ustc.edu.cn, pjzhang@scut.edu.cn

Abstract

Recent studies in deep learning based acoustic echo cancellation proves the benefits of introducing a linear echo cancellation module. However, the convergence problem and potential target speech distortion impose an additional learning burden for the neural network. In this paper, we propose a two-stage progressive neural network consisting of a coarse-stage and a fine-stage module. For the coarse-stage, a light-weighted network module is designed to suppress partial echo and potential noise, where a voice activity detection path is used to enhance the learned features. For the fine-stage, a larger network is employed to deal with the more complex echo path and restore the near-end speech. We have conducted extensive experiments to verify the proposed method, and the results show that the proposed two-stage method provides a superior performance to other state-of-the-art methods.

Index Terms: acoustic echo cancellation, deep learning, two-stage, coarse-stage, fine-stage

1. Introduction

In a hand-free communication system, the acoustic echo is a troublesome distraction and annoying. The echo occurs when a far-end signal gets played out from the loudspeaker and then recorded by the near-end microphone. Acoustic echo cancellation (AEC) aims to suppress the undesired echo picked up by the microphone. There are a wide range of real-world applications where it is extremely desired to remove the echo, such as real time communication [1, 2, 3], smart classroom [4], hand-free system in car [5, 6], and so on.

Considering the importance of the AEC task, numerous studies have been made to address it since the 1990s. Traditional methods [7, 8] try to estimate the echo path with a linear AEC (LAEC) by employing adaptive filter algorithms and then reconstruct the echo to be subtracted from the microphone signal. Varieties of nonlinear processors (NLP) [9, 10, 11] based on digital signal processing are also designed to remove residual echo. However, the results were still far from satisfactory.

More recently, the data-driven AEC models with deep learning (DL) methods have proven to be more powerful [12, 13, 14, 15]. Those methods formulate the AEC as a supervised learning problem in which the mapping function between input signals and the near-end target signal is learned with a deep neural network (DNN). However, the real echo path is extremely complex, which imposes high requirement on the modeling capacity of the DNN. To reduce the modeling burden of the network, most existing DL-based AEC approaches [16, 17, 18] employ a LAEC module based on the adaptive filter algorithms to suppress a large proportion of the linear component of echo. However, there are two disadvantages with the LAEC module: 1) an inappropriate LAEC may cause some near-end speech distortion, and 2) the convergence process makes the performance of the LAEC unstable. Since the LAEC is self-optimizing, the shortcomings of the LAEC would lead to an additional learning burden for the subsequent network.

To better integrate a LAEC module into the DL framework, several differentiable digital signal processing (DDSP) methods [19, 20, 21] have been attempted. Authors in [20] represent the LAEC as a differentiable layer within a DL framework, where a DNN is employed to estimate the step size parameters and reference signal for the LAEC. However, the update process of adaptive filter coefficients still exists in the aforementioned methods, which makes the convergence problem still not effectively resolved. There are also several end-to-end multi-stage DL based methods without using the LAEC module [22, 23, 24]. Authors in [25] proposed a cascade architecture for joint phase and magnitude enhancement. A two-stage structure (*DAEC-64+NRES-64*) [26] decoupling the tasks of echo cancellation and noise suppression is designed, which consists of a deep AEC module and a noise and residual echo suppression module. However, it still remains challenging to balance the echo removal and near-end speech preservation.

To avoid the influence of the LAEC and preserve a better near-end speech quality, this paper explores a novel end-to-end DL based two-stage progressive processing pattern, where the network's capability of modeling complex situations gradually increases. We wish to design a neural network based preprocessing module, which has a similar function as the LAEC that helps to reduce the learning burden of the subsequent model and can be jointly optimized. To this end, a two-stage progressive neural network (TSPNN) consisting of a coarse stage and a fine stage is proposed. At the coarse stage, a light-weighted preliminary AEC module (P-AEC) is first designed to preliminarily suppress partial echo and noise. More importantly, a multi-task learning based near-end voice activity detection (VAD) task is considered, where a VAD penalty loss function is employed to enforce the network to be aware of the near-end speech. After the coarse stage, the signal to echo ratio increases accordingly. However, some residual echo and noise still exist due to the incapability of modeling complex echo scenarios using a light-weighted network structure. At the fine stage, a larger refinement AEC module (R-AEC) is then applied to remove residual echo and noise. Meanwhile, to deal with a potential speech distortion caused by the light-weighted model used in coarse stage, several neighboring time-frequency bins are considered when reconstructing the near-end signal. The coarse stage and fine stage are jointly optimized to avoid the sub-optimization caused by optimizing each stage independently.

*Work conducted when the first author was intern at Bytedance.

10.21437/Interspeech.2023-521

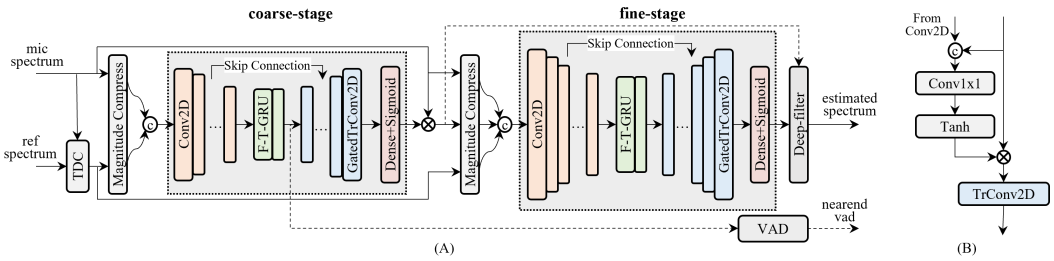

Figure 1: *(A) The flowchart of the proposed two-stage progressive neural network, which consists of a time delay compensation (TDC) block, a coarse-stage module and a fine-stage module. (B) The design of GatedTrConv2D block.*

2. Problem Formulation

In a full-duplex communication system, the microphone signal $y(t)$ can be formulated as

$$y(t) = s(t) + n(t) + h_r(t) * \Gamma(x(t)) \tag{1}$$

where $s(t)$ is the near-end target signal, $n(t)$ is the interference noise, $x(t)$ is the far-end reference signal, $h_r(t)$ is the room impulse response (RIR) representing the echo path, and $\Gamma(*)$ denotes potential non-linear distortion, e.g., loudspeaker distortion and other processing distortion. In general, the AEC task is to remove the echo component $h_r(t) * \Gamma(x(t))$ from $y(t)$. In this paper, we consider a more challenging situation, where the interference noise $n(t)$ should also be suppressed simultaneously.

3. Proposed Method

3.1. Overview

In real communication environments, complex echo scenarios, e.g. extremely low signal-to-echo ratio and heavy reverberation, place extra burden on a single stage model. In this paper, we introduce a two-stage progressive neural network (TSPNN) to perform echo cancellation using an initial coarse processing followed by a final finer level of processing. The two-stage network parameters are jointly optimized to avoid a sub-optimal solution caused by the self-optimization for each stage.

As depicted in Fig. 1 (A), the proposed method consists of a GCC-PHAT [27] based time delay compensator (TDC) for aligning the reference signal and microphone signal, and a neural network based echo canceller: the TSPNN for recovering the near-end clean speech progressively. The TSPNN contains two processing stages: a coarse-stage and a fine-stage. At the coarse-stage, a light-weighted model is first employed to detect the near-end speech and reduce the partial echo and noise. At the fine-stage, a larger model is then designed to suppress the residual echo and noise, and restore the near-end speech.

3.2. Input features

Given a far-end reference signal $x(t) \in \mathbb{R}^{1\times S}$ and a microphone signal $y(t) \in \mathbb{R}^{1\times S}$, the TDC is first applied to predict the aligned reference signal $x'(t) = x(t-\Delta)$, which has been proved to be helpful for the convergence of the AEC model. A short-time Fourier transform (STFT) module is then used to convert $x'(t), y(t)$ into time-frequency domain representations $X, Y \in \mathbb{C}^{F\times T}$ respectively, where F, T denote the number of frequency and time bins respectively. To improve the detection of the low energy near-end signal, a compressing strategy on the raw spectrogram is utilized and the stacked compressed input features $I_{cprs} \in \mathbb{R}^{2\times F\times T}$ are defined as follows.

$$X_{cprs} = |X|^{\alpha} \in \mathbb{R}^{F\times T}, \quad Y_{cprs} = |Y|^{\alpha} \in \mathbb{R}^{F\times T} \tag{2}$$

$$I_{cprs} = stack([X_{cprs}, Y_{cprs}]) \tag{3}$$

where X_{cprs}, Y_{cprs} denote the compressed magnitudes of the far-end reference signal and microphone signal with the compression factor $\alpha \in (0, 1)$, respectively.

3.3. Network structure

3.3.1. Coarse-stage

We employ a light-weighted neural network based preliminary AEC module (P-AEC) at the coarse-stage to remove the partial echo and noise. To enhance the hidden features of near-end clean speech, we employ multi-task learning considering the voice detection and near-end signal reconstruction task simultaneously. The backbone of P-AEC is based on the convolution recurrent network (CRN) [28, 29], which is capable of capturing the local spectral pattern and long-term dependence. Since only a coarse processing is required, a light-weighted CRN is used to reduce the computational cost.

The proposed P-AEC contains five modules: an encoder, a dual-path GRU, a decoder, a mask predictor and a VAD path. The encoder module consists of a series of stacked Conv2D blocks, which captures local spectral features and compresses the frequency dimension to expand the receptive field. Each Conv2D block is a Conv2D layer cascaded with a batch normalization (BN) layer and the parameter rectified linear unit (PReLU) for the activation function. To better model global features over the frequency and time, the dual-path GRU module is applied, which consists of two F-T-GRU layers [30]. Each F-T-GRU layer contains a bi-directional GRU for capturing the harmonic dependence along the frequency dimension and a unidirectional GRU for modeling the long-term dependence along the time dimension. Based on the features extracted by the dual-path GRU module, a decoder module is used to recover the frequency dimension and synthesize the target speech features. The decoder module is symmetrical with the encoder, where the Conv2D layer is replaced with the TransposeConv2D layer. To further filter the features, a gate mechanism is introduced in the decoder module (*GatedTrConv2D*) taking into account the output of the encoder, which is shown in Fig. 1 (B). After the decoder module, the mask prediction module directly predicts a complex ideal ratio mask (cIRM) of the target speech. The mask predictor contains a linear layer and a sigmoid activation function. To enhance the awareness of the near-end clean speech, the VAD path shown in Fig. 1 is designed to predict the

probability of near-end clean speech. The VAD implementation is detailed in Table 1.

Specifically, given an input feature I_{cprs}, the P-AEC predicts a VAD state $P_{vad} \in \mathbb{R}^{2\times T}$ and the cIRM $M_{coarse} \in \mathbb{C}^{F\times T}$ respectively. A coarse estimation of the STFT representation of the target speech $O_{coarse} \in \mathbb{C}^{F\times T}$ is then generated by multiplying the M_{coarse} and the STFT representation of microphone signal Y, where

$$O_{coarse} = Y \otimes M_{coarse} \tag{4}$$

and $\otimes$ denotes the point-wise complex multiplication.

3.3.2. Fine-stage

Since partial echo and noise have been removed at the coarse-stage, the difficulty of echo cancellation and noise suppression has been much reduced. In this case, a larger neural network based refinement AEC module (R-AEC) is employed to remove the residual echo and noise and restore the near-end clean speech.

The backbone of the R-AEC is similar to the P-AEC, which is also based on a CRN and contains four modules: encoder, dual-path GRU, decoder, mask prediction. There are three main differences between the P-AEC and R-AEC in the network structure. Firstly, the network size of the R-AEC is larger, which helps to model the more complex residual echo. Secondly, the VAD path is removed from the R-AEC. Since the echo and noise have been broadly suppressed, the near-end clean speech is easier to be identified by the R-AEC. Thirdly, the R-AEC applies the masking operation in the form of a deep-filter [31] instead of point-wise multiplication. The deep filter based mask strategy takes into account neighboring spectral information and achieves a better harmonic reconstruction.

Similar to the P-AEC, the stacked compressed magnitude $I'_{cprs} \in \mathbb{R}^{3\times F\times T}$ is used as the input feature for the R-AEC, which contains the output of the P-AEC, reference and microphone signals. The R-AEC takes the I'_{cprs} as input and outputs a mask $M_{fine} \in \mathbb{C}^{F\times T\times(2*N_f+1))\times(N_t+N_l+1)}$ in the deep-filter format, where N_f denotes the number of neighbors in each side along the frequency dimension, and N_t, N_l denotes the number of past frames and look-ahead respectively. The estimated mask is then applied directly to the output of the P-AEC, which can be formulated as

$$O_{fine}(f,t) = \sum_{k=-N_f}^{N_f} \sum_{n=-N_t}^{N_l} O_{coarse}(f+k, t+n) \otimes M_{f,t}^{fine}(k,n) \tag{5}$$

where $O_{fine} \in \mathbb{C}^{F\times T}$ is the STFT representation of the estimated target speech. The raw waveform of the target speech is then reconstructed by an inverse-STFT (iSTFT).

3.4. Loss function

We jointly optimize the coarse-stage and fine-stage to avoid the sub-optimization caused by optimizing each stage independently. In our preliminary experiments, we experimentally found that under the two stage schema, the mean absolute error (MAE) loss obtains a higher overall perception score than the power-law compression loss and asymmetrical loss which was shown to be useful in the one stage model. In this case, we optimize the two-stage model with the phase-aware loss based on the MAE loss, which is defined as

$$\mathcal{L}(S,O) = MAE(|S|,|O|) + MAE(S_r,O_r) + MAE(S_i,O_i) \tag{6}$$

$$loss = \omega\mathcal{L}(S, O_{coarse}) + (1-\omega)\mathcal{L}(S, O_{fine}) \tag{7}$$

Table 1: *The configuration of the VAD path.*

Layer	Input size	Ouput size
Conv2D	$32 \times 5 \times T$	$16 \times 5 \times T$
BN+PReLu		
F-GRU	$16 \times 5 \times T$	hidden state:$8 \times 2 \times T$
Reshape	$8 \times 2 \times T$	$16 \times T$
Conv1D	$16 \times T$	$16 \times T$
BN+PReLu		
Conv1D	$16 \times T$	$2 \times T$

*T denotes the time dimension.
*All convolution layers' kernel size and stride are set to 1.

where $S \in \mathbb{C}^{F\times T}$ denotes the groundtruth and $O_{coarse}, O_{fine} \in \mathbb{C}^{F\times T}$ denote the estimated STFT representations of the near-end signal respectively, $|\cdot|$ is the modular operation, S_r, S_i denote the real and imaginary part of the complex number respectively, and the ω is a scalar controlling the degree of focus on each stage. The ω with value of 0.3 is used to soften the optimization of the coarse-stage, which help to alleviate the potential distortion of the near-end clean speech.

To achieve a minimum near-end signal distortion at the coarse stage, the multi-learning strategy is used where a VAD loss is introduced:

$$loss_{vad} = CrossEntropy(P_{vad}, P) \tag{8}$$

where $P_{vad} \in \mathbb{R}^{2\times T}$ is the estimated VAD state and $P \in \mathbb{R}^{1\times T}$ is the groundtruth of the VAD of the near-end speech generated by using the WebRTC-VAD (*https://webrtc.org*). The final loss to be optimized is

$$loss_{final} = loss + \beta loss_{vad} \tag{9}$$

where β is a controlling scalar with value of 0.06.

4. Experiments and Results

4.1. Dataset

We synthesize a total of 1720 h training data by dynamically remixing the clean speech, echo, noise, and RIR sets. Clean speech from the DNS Challenge [32] and LibriSpeech [33] are used as the near-end signal. The echo and reference data pairs are from the far-end single-talk real recording dataset in the AEC Challenge [34, 35]. The noise sets are selected from the DNS Challenge and MUSDB[1]. The RIR set from the DNS Challenge is applied to the clean speech to simulate the reverberation near-end signal. During data mixing, the signal-to-echo ratio (SER), echo-to-noise ratio (ENR) and signal-to-noise ratio (SNR) are set to [-20, 20]dB, [-5, 15]dB and [-5, 15]dB, respectively. All waveforms are resampled at 16kHz.

4.2. Experimental setups

The 20ms Hanning window and 10ms hop-size are adopted to perform the STFT. The model is trained for 60k steps with a batch size of 32 and the speech duration is set to 8s. Adam is used as the optimizer with an initial learning rate of 1e-3. The input magnitude is compressed with the factor $\alpha = 0.3$. In coarse-stage, the number of filters (NOF) of Conv2D layers in the encoder are $\{16, 16, 16, 32, 32, 32, 32, 32\}$ with kernel size $\{[5,1], [1,5], [6,5], [4,3], [6,5], [5,3], [3,5], [3,3]\}$ and stride $\{[1,1], [1,1], [2,1], [2,1], [2,1], [2,1], [2,1], [1,1]\}$, and the hidden size of F-T-GRUs in the dual-path GRU is

[1]https://doi.org/10.5281/zenodo.1117372

Table 2: *Ablation study for two-stage model on a real testset.*

	single-talk		double-talk		
	FE	NE	echo	deg	avg
P_{vad}-R_{df}-*AEC*	4.76	**4.12**	4.68	**4.14**	**4.425**
P-R_{df}-*AEC*	4.73	4.12	**4.69**	4.10	4.410
P-$R_{vad,df}$-*AEC*	**4.77**	4.09	4.65	4.02	4.383
R_{vad}-P_{df}-*AEC*	4.76	4.10	4.68	4.11	4.413

Table 3: *Comparison with single-stage models on the blind test [34]. AECMOS denotes the average score on all scenarios.*

	AECMOS	#param (M)	complexity (G Macs)	data (h)
$R_{df,large}$-*AEC*	4.439	0.94	2.98	700
LAEC-$R_{df,large}$-*AEC*	4.434	0.94	2.98	700
$R_{df,large}$-*AEC Plus*	4.442	2.45	7.23	700
P_{vad}-R_{df}-*AEC*(**ours**)	4.506	1.27	2.82	700
P_{vad}-R_{df}-*AEC*(**ours**)	**4.525**	1.27	2.82	1720

$\{[32, 64], [32, 32]\}$. In the fine-stage, the configuration of the network is similar to the coarse-stage, where the NOF of Conv2D layers is $\{16, 16, 32, 32, 64, 64, 64, 64\}$ and the hidden size in the dual-path GRU is $\{[64, 128], [64, 64]\}$. The N_f, N_t and N_l are set to 3, 3, and 1 respectively when performing the deep filtering. All network layers are configured as causal and the BN is only performed during training.

The performance of the proposed method is evaluated using the AECMOS model [36], which can be used to estimate the mean opinion score (MOS) with great accuracy. The AECMOS model predicts an *echo degradation* MOS_e for judging the degradation from the echo and an *other degradation* MOS_o for judging other degradations (such as noise, missing audio, distortions and cut-outs). For far-end single-talk (FE), the MOS_e is applied. For near-end single-talk (NE), the MOS_o is applied. For double-talk, both MOS_e and MOS_o are applied, which are denoted as *echo* and *deg* respectively. The predicted MOS scores are from 1 to 5. A higher score corresponds to a constructed signal with better quality.

4.3. Results and analysis

To validate the configuration of the proposed method, the ablation experiments are first conducted and the results are shown in Table 2. The proposed method is denoted as ***P_{vad}-R_{df}-AEC***. After removing the VAD path from the coarse-stage (***P-R_{df}-AEC***), the MOS_o is degraded by 0.04 in the double-talk scenario. Moving the VAD path from the coarse-stage to the fine-stage (***P-$R_{vad,df}$-AEC***) also leads to a degradation on the AECMOS. In addition, switching the CRN configurations between the coarse-stage and fine-stage (***R_{vad}-P_{df}-AEC***) also results in a performance degradation. The schema of "small coarse-stage, large fine-stage" improve the upper bound of the performance, as it has the ability to recover the clean speech gradually.

To further verify the feasibility of the proposed two-stage strategy, we compare it with three large single-stage baselines constructed by removing the coarse-stage and enlarging the network in fine-stage: ***$R_{df,large}$-AEC***: increasing the N_f and N_t in deep filter to 5 and 5 respectively; ***$R_{df,large}$-AEC Plus***: further increasing the NOF of Conv2D layers to $\{32, 32, 64, 64, 64, 96, 96, 128\}$ and the hidden size in the dual-path GRU to $\{[128, 256], [128, 128]\}$; and ***LAEC-$R_{df,large}$-AEC***: the ***$R_{df,large}$-AEC*** with the **LAEC**(a subband cross-correlation based linear AEC with NLMS) output as network input. As seen in Table 3, the proposed two-stage method outperforms the one-stage methods with less computational cost, and simply introducing the LAEC as a pre-processor limits the performance which is consistent with the conclusion in [16].

[2]https://www.speex.org/

Table 4: *Comparison with other methods on the blind test [34].*

	single-talk		double-talk		
	FE	NE	echo	deg	avg
Unprocessed	1.86	4.07	2.14	4.14	3.053
Speex-AEC[2]	3.42	4.08	2.78	3.81	3.523
LAEC	3.36	4.14	2.92	3.94	3.590
coarse-stage	4.79	4.27	4.55	4.01	4.405
rank 1 *	*4.59*	*4.25*	*4.69*	*4.18*	*4.428*
DAEC-64+NRES-64[26]	4.35	4.18	4.55	4.25	4.333
P_{vad}-R_{df}-*AEC*(**ours**)	**4.80**	**4.32**	**4.69**	**4.29**	**4.525**

*The MOS of rank 1 is used from Challenge[34]. The AECMOS model estimates MOS with high accuracy[36].

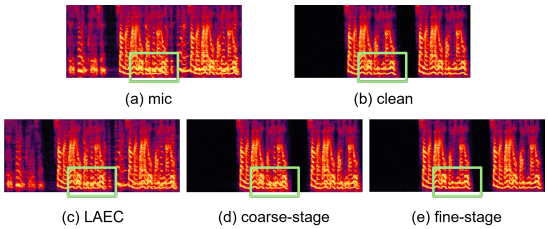

Figure 2: *Visualization of the results on a double-talk scenario.*

To understand the effect of the proposed method, we also compare the difference between the outputs of the coarse-stage and LAEC. Table 4 shows a overall comparison, from which we can see that the neural network based pre-processing is significantly superior to the LAEC. In addition, Fig. 2 compares the coarse-stage with LAEC on a double-talk scenario. It's obvious that the output of the coarse-stage has fewer residual echo, which facilitates the target speech restoration for the fine-stage. The fine-stage further reconstructs the near-end signal quite cleanly.

To validate the effectiveness of the proposed method, we also compare with other state-of-the-art methods on the 2nd AEC Challenge dataset[34] in Table 4. Our proposed method outperforms the first ranked system by a large margin on all scenarios, which proves the effectiveness of the proposed two-stage method. It's worth noting that *DAEC-64+NRES-64* is also a two stage method. Our proposed method achieves a higher quality restoration of the near-end speech compared to *DAEC-64+NRES-64*, which demonstrates the proposed two-stage pattern is more powerful and effective. The model details and results can be found from *https://github.com/enhancer12/TSPNN*.

5. Conclusion

In this paper, we proposed a novel two-stage progressive neural network for acoustic echo cancellation with a modeling capacity from low-quality to high-quality. The experimental results shows that the two-stage strategy helps to improve performance and the improvement doesn't come at the expense of an increase in model size. The proposed method has been shown to be superior to other state-of-the-art methods.

6. References

[1] C. Zheng, Y. Zhou, X. Peng, Y. Zhang, and Y. Lu, "Time-variance aware dynamic kernel generation for real-time acoustic echo cancellation," *IEEE Signal Processing Letters*, vol. 29, pp. 967–971, 2022.

[2] X. Wang, S. Zhu, and L. Zhao, "A real-time dereverberation algorithm combined with echo cancellation," in *ICMTMA*. IEEE, 2020, pp. 464–467.

[3] L. Shen and Z. Long, "Principle and algorithms of webrtc echo cancellation," in *The 2nd International Conference on Computing and Data Science*, 2021, pp. 1–5.

[4] R. Guntha, B. Hariharan, and P. V. Rangan, "Analysis of echo cancellation techniques in multi-perspective smart classroom," in *ICACCI*. IEEE, 2016, pp. 1135–1140.

[5] J. Franzen and T. Fingscheidt, "A delay-flexible stereo acoustic echo cancellation for dft-based in-car communication (icc) systems." in *Interspeech*, 2017, pp. 181–185.

[6] J. Franzen, I. M. zum Alten Borgloh, and T. Fingscheidt, "On the benefit of a stereo acoustic echo cancellation in an in-car communication system," in *Speech Communication; 13th ITG-Symposium*. VDE, 2018, pp. 1–5.

[7] A. Gilloire and M. Vetterli, "Adaptive filtering in sub-bands with critical sampling: analysis, experiments, and application to acoustic echo cancellation," *IEEE transactions on signal processing*, vol. 40, no. ARTICLE, pp. 1862–1875, 1992.

[8] C. Paleologu, S. Ciochină, J. Benesty, and S. L. Grant, "An overview on optimized nlms algorithms for acoustic echo cancellation," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, pp. 1–19, 2015.

[9] O. Tanrikulu and K. Dogançay, "A new non-linear processor (nlp) for background continuity in echo control," in *ICASSP*, vol. 5. IEEE, 2003, pp. V–588.

[10] M. I. Mossi, C. Yemdji, N. Evans, C. Beaugeant, and P. Degry, "Robust and low-cost cascaded non-linear acoustic echo cancellation," in *ICASSP*. IEEE, 2011, pp. 89–92.

[11] M. Z. Ikram, "Non-linear acoustic echo cancellation using cascaded kalman filtering," in *ICASSP*. IEEE, 2014, pp. 1320–1324.

[12] Q. Lei, H. Chen, J. Hou, L. Chen, and L. Dai, "Deep neural network based regression approach for acoustic echo cancellation," in *ICMSSP*, 2019, pp. 94–98.

[13] Y.-C. Tsai, K.-W. Liang, and P.-C. Chang, "Acoustic echo cancellation based on recurrent neural network," in *APSIPA ASC*. IEEE, 2020, pp. 88–91.

[14] Y. Zhang, C. Deng, S. Ma, Y. Sha, H. Song, and X. Li, "Generative adversarial network based acoustic echo cancellation." in *Interspeech*, 2020, pp. 3945–3949.

[15] K. N. Watcharasupat, T. N. T. Nguyen, W.-S. Gan, S. Zhao, and B. Ma, "End-to-end complex-valued multidilated convolutional neural network for joint acoustic echo cancellation and noise suppression," in *ICASSP*. IEEE, 2022, pp. 656–660.

[16] G. Zhang, L. Yu, C. Wang, and J. Wei, "Multi-scale temporal frequency convolutional network with axial attention for speech enhancement," in *ICASSP*. IEEE, 2022, pp. 9122–9126.

[17] H. Zhao, N. Li, R. Han, L. Chen, X. Zheng, C. Zhang, L. Guo, and B. Yu, "A deep hierarchical fusion network for fullband acoustic echo cancellation," in *ICASSP*. IEEE, 2022, pp. 9112–9116.

[18] S. Zhang, Z. Wang, J. Sun, Y. Fu, B. Tian, Q. Fu, and L. Xie, "Multi-task deep residual echo suppression with echo-aware loss," in *ICASSP*. IEEE, 2022, pp. 9127–9131.

[19] J. Engel, L. H. Hantrakul, C. Gu, and A. Roberts, "Ddsp: Differentiable digital signal processing," in *ICLR*, 2020. [Online]. Available: https://openreview.net/forum?id=B1x1ma4tDr

[20] H. Zhang, S. Kandadai, H. Rao, M. Kim, T. Pruthi, and T. Kristjansson, "Deep adaptive aec: Hybrid of deep learning and adaptive acoustic echo cancellation," in *ICASSP*. IEEE, 2022, pp. 756–760.

[21] M. Kawamura, T. Nakamura, D. Kitamura, H. Saruwatari, Y. Takahashi, and K. Kondo, "Differentiable digital signal processing mixture model for synthesis parameter extraction from mixture of harmonic sounds," in *ICASSP*. IEEE, 2022, pp. 941–945.

[22] M. Strake, B. Defraene, K. Fluyt, W. Tirry, and T. Fingscheidt, "Separated noise suppression and speech restoration: Lstm-based speech enhancement in two stages," in *WASPAA*. IEEE, 2019, pp. 239–243.

[23] M. Strake, B. Defraene, K. Fluyt, W. Tirry, and T. Fingscheidt, "Speech enhancement by lstm-based noise suppression followed by cnn-based speech restoration," *EURASIP Journal on Advances in Signal Processing*, vol. 2020, pp. 1–26, 2020.

[24] A. Li, W. Liu, C. Zheng, C. Fan, and X. Li, "Two heads are better than one: A two-stage complex spectral mapping approach for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1829–1843, 2021.

[25] H. Zhang and D. Wang, "Neural cascade architecture for joint acoustic echo and noise suppression," in *ICASSP*. IEEE, 2022, pp. 671–675.

[26] S. Braun and M. L. Valero, "Task splitting for dnn-based acoustic echo and noise removal," in *IWAENC*. IEEE, 2022, pp. 1–5.

[27] 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.

[28] X. Le, H. Chen, K. Chen, and J. Lu, "DPCRN: Dual-Path Convolution Recurrent Network for Single Channel Speech Enhancement," in *Proc. Interspeech 2021*, 2021, pp. 2811–2815.

[29] F. Dang, H. Chen, and P. Zhang, "Dpt-fsnet: Dual-path transformer based full-band and sub-band fusion network for speech enhancement," in *ICASSP*. IEEE, 2022, pp. 6857–6861.

[30] X. Le, L. Chen, C. He, Y. Guo, C. Chen, X. Xia, and J. Lu, "Personalized speech enhancement combining band-split rnn and speaker attentive module," in *ICASSP*. IEEE, 2023, pp. 1–2.

[31] H. Schroter, A. N. Escalante-B, T. Rosenkranz, and A. Maier, "Deepfilternet: A low complexity speech enhancement framework for full-band audio based on deep filtering," in *ICASSP*. IEEE, 2022, pp. 7407–7411.

[32] H. Dubey, V. Gopal, R. Cutler, A. Aazami, S. Matusevych, S. Braun, S. E. Eskimez, M. Thakker, T. Yoshioka, H. Gamper *et al.*, "Icassp 2022 deep noise suppression challenge," in *ICASSP*. IEEE, 2022, pp. 9271–9275.

[33] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *ICASSP*. IEEE, 2015, pp. 5206–5210.

[34] R. Cutler, A. Saabas, T. Pärnamaa, M. Loide, S. Sootla, M. Purin, H. Gamper, S. Braun, K. Sørensen, R. Aichner *et al.*, "Interspeech 2021 acoustic echo cancellation challenge." in *Interspeech*, 2021, pp. 4748–4752.

[35] R. Cutler, A. Saabas, T. Parnamaa, M. Purin, H. Gamper, S. Braun, K. Sørensen, and R. Aichner, "Icassp 2022 acoustic echo cancellation challenge," in *ICASSP*. IEEE, 2022, pp. 9107–9111.

[36] M. Purin, S. Sootla, M. Sponza, A. Saabas, and R. Cutler, "Aecmos: A speech quality assessment metric for echo impairment," in *ICASSP*. IEEE, 2022, pp. 901–905.