Interspeech 2024
1-5 September 2024, Kos, Greece

Multi-mic Echo Cancellation Coalesced with Beamforming for Real World Adverse Acoustic Conditions

Premanand Nayak, Kamini Sabu, M. Ali Basha Shaik

Language AI R&D,
Samsung Research Institute-Bangalore, India
{premanand.n, kamini.sabu, m.shaik}@samsung.com

Abstract

Robust acoustic echo cancellation (AEC) is essential for voice enabled smart devices. Multi-channel signals are used in AEC along with beamformer (BF) for better residual echo suppression (RES). In this work, we introduce a deep neural network (DNN) based novel unified framework for multi-microphone AEC (MMAEC) and RES under adverse signal-to-echo (SER) conditions. We propose the use of deep-MVDR which uses deep steering vector and deep power spectral density (Deep PSD) estimator to conceptually implement minimum variance distortionless beamformer. We also introduce additional novelty in our framework by jointly training the MMAEC and deep-MVDR modules. Both of these methods give consistent significant improvement in ERLE which is further enriched by the incorporation of playback reconstruction loss. Our system outperforms competitive baselines while being robust in adverse real-world conditions such as very low input SER, dominant far-end sources, and moving near-end speech sources.

Index Terms: multi-mic, real-world audio, unified, joint-training

1. Introduction

In today's smart home era, accurate detection of user's command for devices like voice assistants, TVs and soundbars etc., is still an open problem due to dominant playback content and very low SERs. As the playback signal is known a-priori, standard AEC methods can be used to cancel the playback signal. However, AEC task becomes challenging when both playback and target/nearend speech are active at the same time, more so ever when the playback is more dominant. Also, the loudspeaker's playback signal undergoes several non-linear transformations. Moreover, the movement of near-end speech source (i.e., moving user) poses an additional challenge. Conventional filter methods [1, 2, 3, 4, 5] applied to capture the room impulse response (RIR) for a mic and loudspeaker position using error minimization techniques. However, they fail to capture the non-linear distortions introduced in the playback. Therefore, DNNs which inherently capture non-linearities at the playback, were used in [6, 7]for residual echo suppressed post these filter based systems. Although they are able to suppress the residual echo significantly, playback non-linear distortion's are not completely nullified. Recently, neural networks have been employed to estimate either the RIR filter or the target echo canceled signal. These mostly rely on Short Time Fourier Transform (STFT) mask estimations [8, 9, 10, 11]. along with neural network fusion in [12, 13]. Although they do not need explicit post-processing, but, they still fall short in handling real-world challenging issues such as dominant playback in double-talk conditions, moving sources, very low SER conditions etc.

2. Prior Art and Novelty

Multiple mics usage helps to estimate source direction from AEC, leading to efficient echo suppression. Conventional methods estimate channel-wise RIR filter to provide the required multi-channel AEC output [14, 15, 16]. Neural network (NN) encoder-decoder methods followed by beamformer generally outperform these [17, 18, 19, 20]. The NN combined systems work well for moving near-end source. But, the performance degrades in the presence of a dominant speech playback.

We make an attempt to address these complex issues in multi-mic AEC for adverse SER and both static and moving near-end sources. We propose a novel two-stage architecture: The first stage MMAEC performs multi-channel first pass echo cancellation, while in the second stage, D-MVDR performs RES by generating a robust beamforming weight vector. This effectively acts like a latent variable in our framework. The two stages are jointly trained using combined reconstruction objectives: (a) single channel near-end speech (b) echo canceled mic signals consisting of only target speech along with (a). We also show that the ancillary objective of playback signal reconstruction further aids in better estimation of the target source. To the best of authors' knowledge, this is one of the earliest work proposing such a kind of coalesced approach for enhancing multi-microphone echo cancellation performance. Here, we target real-world audios under adverse low SERs with complex conditions like static and moving source. We pragmatically validate and provide the incremental evidence by comparing with competitive state-of-the-art (SoTA) approaches using ERLE and PESQ metrics.

3. Problem Formulation

Mathematically, an N-channel microphone signal in the presence of playback signal at time instant n can be modelled as: $\mathbf{m}[n] = \mathbf{r}[n] + \mathbf{d}[n] + \mathbf{b}[n] \in \mathbb{R}^N$ Here, $\mathbf{r}[n] = \mathbf{h}_s[n] \circledast s[n]$, $\mathbf{d}[n] = \mathbf{h}_x[n] \circledast f_l(x[n])$, and $\mathbf{b}[n]$ represent the copies of near-end user speech signal, playback signal, and background noise signal components received at the microphones, respectively. Here, $s[n]$: speech signal (near-end), and $\mathbf{h}_s[n]$: RIR for the microphone-source pairs, $x[n]$: loudspeaker's reference playback signal, $f_l(.)$: nonlinear transformations introduced by loudspeaker device, and $\mathbf{h}_x[n]$: RIR for the loudspeaker-microphone pairs. After applying short time Fourier transform (STFT), the microphone signal at time t and frequency f is given by:

$$\mathbf{M}(t,f) = \mathbf{D}(t,f) + \mathbf{R}(t,f) + \mathbf{B}(t,f) \in \mathbb{C}^N \quad (1)$$

The aim of this work is to suppress $\mathbf{d}[n]$ and $\mathbf{b}[n]$ while retaining $\hat{s}[n]$ such that $\hat{s}[n] \mapsto s[n]$. We propose MMAEC for acoustic echo suppression, where beamformer operates as residual echo suppressor. Let θ_{aec} and θ_{BF} represent the AEC and

10.21437/Interspeech.2024-834

BF model parameters, respectively. Then our proposed system performs the composite transformation $F(.) = F_{\theta_{aec}.\theta_{BF}} = F(\theta_{AEC}) \circ F(\theta_{BF})$ on microphone and playback, $\mathbf{M}(t,f)$ and $X(t,f)$ respectively, such that:

$$\hat{s}[n] = ISTFT(F(\mathbf{M}(t,f), X(t,f))) \quad (2)$$

4. Proposed Approach

Fig. 1 shows the proposed architecture of our system. MMAEC uses multi-mic and playback signal STFT, $\mathbf{M}$ and X respectively, to estimate echo-canceled signal $\hat{\mathbf{M}}$. D-MVDR performs steering vector (SV) estimation and RES of $\hat{\mathbf{M}}$ to obtain BF-MVDR weight vector. Then, target speech is the inverse STFT of BF output. Elaborate details are delineated in the following sections.

4.1. Deep Neural MMAEC

MMAEC uses convolutional recurrent network (CRN) architecture as shown in Fig. 2. The real and imaginary components of $\mathbf{M}$ and X concatenated across the channel dimension form the 2 (N+1) channel input to the model. The pair-wise identical encoder and decoder blocks use 2D convolution filters with frequency strides to capture time-frequency information. The encoder layers successively use more filters while compressing the frequency dimension. Alternatively, the successive decoder layers up-sample the frequency information for reconstruction while decreasing the filters. The encoder layers are followed by two GRU layers, T-GRU and F-GRU. T-GRU sequential axis is the initial time dimension while F-GRU's axis learns across the frequency dimension. T-GRU captures the temporal sequence information across all channels, whereas, F-GRU captures the information dependencies across frequency bins. We add skip connections between encoder and decoder layers to retain the relevant information gradients as shown in Fig. 2. The MMAEC output is a channel-level real-imaginary mask, $\mathbf{Z}(t,f) \in \mathbb{R}^{2N}$. Its element-wise multiplication with the respective microphone input $\mathbf{M}(t,f)$ produces the first stage echo canceled signal $\hat{\mathbf{M}}(t,f)$.

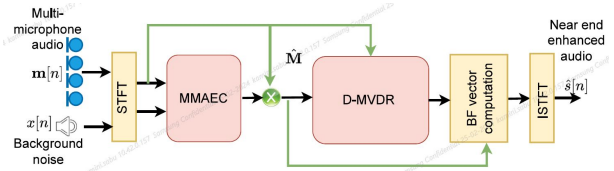

Figure 1: *Schematic diagram depicting the proposed system*

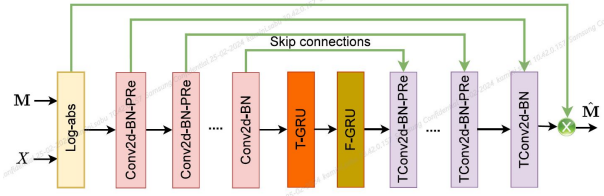

Figure 2: *MMAEC Neural Architecture*

4.2. D-MVDR

The proposed D-MVDR module comprises of Deep PSD and Deep SV modules as shown in Fig. 3. Deep PSD block estimates a second pass of echo suppressed signal $\mathbf{P}$ from $\hat{\mathbf{M}}$, while Deep SV produces the steering estimates for the target/near-end signal and the playback signal. These combined with $\hat{\mathbf{M}}(t,f)$ and $\mathbf{M}(t,f)$ form the essential components for the MVDR weight vector for both target/near-end and playback. The MVDR weight vector is represented in terms of PSD matrix $\boldsymbol{\phi}_I(f)$ and SV $\mathbf{v}(t,f)$ as:

$$\hat{\mathbf{w}}(t,f) = \frac{\boldsymbol{\phi}_I^{-1}(f)\mathbf{v}(t,f)}{\mathbf{v}(t,f)^H \boldsymbol{\phi}_I^{-1}(f)\mathbf{v}(t,f)} \quad (3)$$

Deep PSD consists of convolutional encoder-decoder blocks, whereas the bottleneck layer uses only single-layer LSTM acting along time dimension. Deep PSD output is complex STFT $\mathbf{P}(t,f)$, a suppressed echo signal estimate of $\hat{\mathbf{M}}(t,f)$. It is subtracted from $\hat{\mathbf{M}}(t,f)$ to obtain the interference signal estimate $\mathbf{I}(t,f)$ (see Fig. 3). PSD matrix $\boldsymbol{\phi}_I(f)$ obtained as a moving sum of the rank-1 matrices across past T frames, as shown in Eq. 4.

$$\boldsymbol{\phi}_I(f) = \frac{1}{T}\sum^{T} \mathbf{I}(t,f)\mathbf{I}^H(t,f) - \bar{I}(t,f) \quad (4)$$

Similarly, speech PSD matrix $\boldsymbol{\phi}_P(f)$ is computed for $\mathbf{P}(t,f)$.

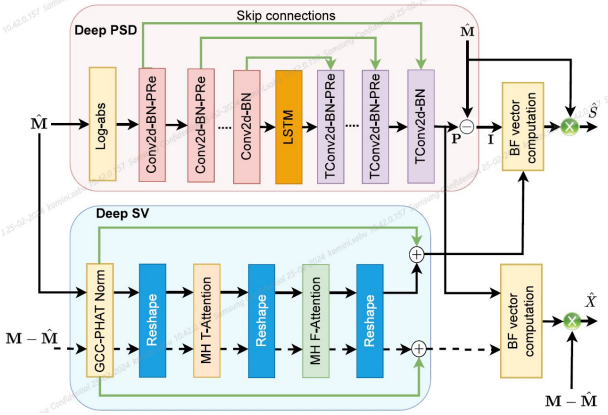

Figure 3: *Proposed D-MVDR Beamforming Architecture*

For SV computation, Deep SV uses time-frequency bin-wise GCC normalization called GCC-Phase transform (GCC-PHAT) [21] to get channel-wise features given by (channel 1 used as reference without any loss of generality):

$$\mathbf{U}_i(t,f) = \frac{\mathbf{C}_1(t,f)\mathbf{C}_i^*(t,f)}{||\mathbf{C}_1(t,f)||\,||\mathbf{C}_i^*(t,f)||} \forall i \neq 1 \quad (5)$$

Here, $\mathbf{C}_i$s represent either the MMAEC processed output $\hat{\mathbf{M}}(t,f)$ or the residual interference signal $\mathbf{M}(t,f)$-$\hat{\mathbf{M}}(t,f)$, both used parallely in the Deep SV module. Two-head complex (one block of magnitude and one for phase) Time Attention and Frequency Attention modules then learn the attention masks for GCC-PHAT transform features across time and then across frequency. This produces complex steering estimates $\mathbf{v}_s(t,f)$ and $\mathbf{v}_x(t,f)$, respectively. Eq. 3 then gives the BF weight vector $\mathbf{w}_s(f)$ for near-end speech using $\{\mathbf{v}_s(t,f), \boldsymbol{\phi}_x(f)\}$ pair and thus near-end signal STFT can be given as:

$$\hat{S}(t,f) = \mathbf{w}_s^H(t,f)\hat{\mathbf{M}}(t,f) \quad (6)$$

Similarly, the far-end signal $\hat{X}(t,f)$ is obtained by multiplying interference signal ($\mathbf{M}(t,f)$ - $\hat{\mathbf{M}}(t,f)$) with BF weight vector $\mathbf{w}_x(f)$ estimated using $\{\mathbf{v}_x(t,f), \boldsymbol{\phi}_P(f)\}$ pair.

4.3. Joint Training Objective Function

We perform joint training of MMAEC and D-MVDR using multi-reconstruction objectives. We use L_{BF} referring to mean squared error (MSE) between the target $\mathbf{S}$ and the estimated clean speech $\hat{\mathbf{S}}$. We investigate the performance of MMAEC model using another reconstruction loss L_{aec} - MSE between MMAEC multi-channel output $\hat{\mathbf{M}}$ and the multi-channel target $\mathbf{M}_{near} \in \mathbb{C}^N$ (generated at microphone in the absence of playback). Joint training loss function L_{joint} is shown in Eq. 7. Similarly, multi-task training loss L_{total}, that includes L_{res} i.e. MSE loss between the playback X and the estimated signal $\hat{X}$ is also added, as shown in Eq. 8. λ_a and λ_b are selected through hyper-parameter tuning.

$$L_{joint} = \lambda_a L_{aec} + (1 - \lambda_a) L_{BF}; \quad (7)$$

$$L_{total} = \lambda_b L_{joint} + (1 - \lambda_b) L_{res}; \quad (8)$$

Table 1: *Single-channel AEC results (**stationary** near-end source test set, Loss: L_{aec}*

Model Type	PESQ	ERLE (dB)
NLMS [1]	2.7	5.4
CRN [8]	3.1	7.2
FTLSTM [13]	3.1	9.3
MMAEC (Proposed)	**3.6**	**11.7**

5. Datasets

We use three datasets in this work. We use opensource MS-SNSD [22] and AEC-Challenge [23] datasets to simulate (with Pyroomacoustics [24]) multi-channel noisy audio with near-end source and playback source chosen from clean subset. Noise subset spans various indoor and outdoor scenarios including babble noise. Playback and near-end speech use 10-20 sec long utterances from clean subset. A linear 4-channel microphone array configuration with uniform spacing of 10 cm or a circular microphone array with a diameter of 10 cm is used. Near-end speech signal source is located in the far-field at varied distances of 1-3 meters (with variable room size of : 3m-6m x 4m-6m). Reverberation time (RT60) varies between 0.3-0.6s. Other non-linear distortions like loudspeaker distortions are implemented as per [17]. We use $\sim$85 hours of 16KHz multi-mic audio with playback at input (0 dB to -15 dB) SER. 20 hrs of data has near-end source moving in linear steps (positions changed every 0.5sec). Microphone and loudspeaker playback position remains constant. So impulse response of echo did not vary, hence no echo path change was considered for near-end source movement. However, near-end source moves around this mic setup (180 degree) piecewise-linear fashion, mimicking real-world movement cases. We ensure each recording contains approx. equi-proportions single-talk near-end, far-end and doubletalk regions. We also generate a corresponding audio without playback to serve as a target for MMAEC training for every simulated recording. Total of 10 hrs of this is used for validation. The simulated test set contains $\sim$4 hrs of stationary recordings and 1hr of moving speaker recordings.

Our real-world dataset comprise of $\sim$10 hrs of smart TV stereo recordings with a 1-ch playback. Librispeech [25] dev set is considered as a near-end source, while the controlled SER playback comes from TV internal speakers playing US English news content. The near-end source in the far-field is stationary at a distance of 1.5m to 2.5m from TV. A tether is connected to the TV to acquire the microphone and playback signals. A total of 5 hrs of this data is used for training and validation, while 5 hrs is used as testset. We plan to release this dataset along with benchmarked model results. This could act as a platform to evaluate and further enhance AEC performance on real-world ambience variational audios.

6. Experiments and Results

We use 512-point STFTs as input with a hamming window size of 20 ms and 50% hop computed for every microphone channel, for our proposed MMAEC+D-MVDR system and state-of-the-art (SoTA) system comparisons (except the time-domain filter estimation methods). The real and imaginary parts of the complex STFT are concatenated across the channel dimension before 2D convolution.

All encoder layers use Conv2d and decoder layers use Transpose Conv2d with (4,4) kernel and (2,4) stride across (F,T) dimensions with padding across T dimension. The number of Conv2d filters are doubled at every successive layer. The Conv layers are followed by batch normalization and PRelu activation. The LSTM/T-GRU/F-GRU feature size depends on the retained frequency/time dimension and the number of kernels used in the final encoder layer, 256 and 4 for both MMAEC and Deep PSD, respectively. PSD matrix for 'T' frames is computed over every 0.3s.We use Adam optimizer with an exponential learning rate scheduler starting at an initial learning rate of 0.005. All the models are trained over 50 epochs. All trainings were done using an in-house NVIDIA A100 40GB GPU, with a total training convergence duration of $\sim$48 hrs. Proposed and compared neural network approaches model paramaters are around 0.2M in size.

6.1. Baseline and State-of-the-Art Systems

We compare our proposed approach with various SoTA multi-mic AEC techniques: (a) adaptive signal processing based NLMS [1, 26], (b) deep neural CRN [17], and (c) FTLSTM model [13]. As these are all single channel methods, we use only one microphone channel for performance comparison. We also train their multi-microphone enhancement counterparts, where single-channel AEC is applied across each channel followed by MVDR beamforming. BF can be signal-processing based (S-MVDR) used in [17] or deep MVDR (D-MVDR). We did analyze other beamforming approaches like [16] but chose to compare with S-MVDR as there wasn't any significant improvement in AEC RES with those. In D-MVDR case, neural network SoTA or MMAEC is trained jointly with D-MVDR. We compare these SoTA with our proposed MMAEC and the jointly trained MMAEC+DMVDR approach (with L_{BF} or L_{joint}). In addition, we train the SoTA methods and our proposed system with playback reconstruction loss (L_{total}). All neural network first-stage echo cancellation systems have roughly 2M parameters and D-MVDR alone has 0.5M parameters. Algorithmic latency for the end-to-end output is 15ms. We compare all the methods using Echo Return Loss Enhancement (ERLE) and Perceptual Evaluation Of Speech Quality (PESQ) measures.

6.2. Results and Discussion

As a fundamental approach, we train the baseline methods for single channel AEC using L_{aec} loss. Only the first channel of input microphone signal is used to get the results shown in Ta-

Table 2: *Results comparisons (**stationary** near-end source test set)*

Model Type	S-MVDR Beamformer PESQ L_{BF}	L_{joint}	L_{total}	ERLE (dB) L_{BF}	L_{joint}	L_{total}	D-MVDR Beamformer PESQ L_{BF}	L_{joint}	L_{total}	ERLE (dB) L_{BF}	L_{joint}	L_{total}
NLMS [1]	3.0	2.9	3.0	5.9	7.5	10.2	3.1	3.2	3.4	7.5	10.5	13.6
CRN [17]	3.3	3.3	3.2	9.2	9.7	14.7	3.1	3.4	3.5	11.3	15.6	19.8
FTLSTM [13]	3.4	3.3	3.4	12.6	13.1	15.6	3.1	3.2	3.3	15.1	16.9	19.4
MMAEC (Prop.)	3.6	3.7	**3.7**	12.3	16.9	**18.2**	3.7	3.7	**3.9**	14.2	19.8	**24.3**

Table 3: *Results ((**moving near-end** source), **simulated** Vs **real world** test set (TV recordings). Loss: L_{total} loss*

Model Type	BF	Simulated PESQ	Simulated ERLE (dB)	Real world PESQ	Real world ERLE (dB)
NLMS [1]	S	3.2	9.2	3.6	6.9
CRN [17]	S	3.7	11.5	3.7	8.6
FTLSTM [13]	S	3.9	18.7	3.9	9.1
MMAEC (Proposed)	S	3.5	17.3	**3.9**	**10.1**
NLMS [1]	D	3.1	12.4	3.6	7.1
CRN [17]	D	3.6	17.6	3.6	7.9
FTLSTM [13]	D	3.7	19.3	3.5	10.3
MMAEC (Proposed)	D	3.9	18.2	**4.2**	**12.4**

Table 4: *ERLE (dB) Results - Adverse SERs with L_{total} loss*

Model Type	SER = -10dB S BF	SER = -10dB D BF	SER = -15dB S BF	SER = -15dB D BF
NLMS [1]	3.6	6.2	3.6	5.8
CRN [17]	12.7	17.5	10.1	10.9
FTLSTM [13]	16.7	16.2	11.5	13.1
MMAEC (Proposed)	14.9	**20.3**	12.7	**17.8**

ble 1. Our MMAEC module trained with near-end only data for L_{aec} loss sees 2dB improvement over the best baseline method with a significant improvement in PESQ. This shows the effectiveness of training multi-mic systems. We further perform joint training of first pass AEC block and second pass RES block (multi-channel MVDR BF). As discussed in Section 6.1, we apply single-channel AEC on input channel-wise followed by beamforming. Table 2 compares the performance of proposed MMAEC system and multi-channel versions of SoTA single-channel methods with RES using S-MVDR and D-MVDR. We observe consistent and noticable improvements in ERLE from 12.3dB to 14.2dB along with boost in PESQ scores as shown in Table 1. Further breaking it down concerning joint training objectives, L_{total} and L_{joint} are performing better for all methods, showing an improvement in PESQ and ERLE from 18.2dB with S-MVDR to 24.3dB for D-MVDR. The loss function weights λ_a and λ_b (Eq. 8) take values 0.39 and 0.76 respectively after hyper-parameter tuning. This indicates that the basic LBF loss gets almost 50% weightage while remaining 50% is almost equally divided between AEC module learning and playback reconstruction. We observe that both L_{aec} loss and playback reconstruction loss L_{res} play very important role in improving the near-end signal reconstruction for both RES, as shown in Table 2.

We demonstrate the performance of AEC+RES methods for moving source in Table 3. We see that D-MVDR performs consistently better than S-MVDR though both perceptually fare equally well. The Deep SV based estimation of dual steering vectors (each for near-end and far-end source) in D-MVDR can be considered as the key factor responsible for the improvement. Among the single-channel front-end methods, FTLSTM shows overall better PESQ and ERLE performance on this simulated dataset. Observations with real testset also show significant improvement of 2.3dB with MMAEC+D-DMVDR. We also observed the best perceptual quality with MMAEC+D-MVDR enhancement as we listened to the enhanced audio from this testset, especially for high TV volume.

Finally, we compare the systems for harsh playback conditions, particularly SERs -10dB and -15dB in Table 4. MMAEC+D-MVDR outperforms others, with a significant gain of 4dB for -10dB SER and 4.7dB for -15dB SER. The other AEC single-channel methods also show improvement when combined with D-MVDR. The proposed architecture outperforms FTLSTM model under very low SER conditions as can be seen from Table 4.

7. Conclusions

We introduced a novel coalesced joint-training framework for multi-microphone AEC and residual echo suppression. We validated our approach as well on various state-of-the-art single and multi-microphone systems under adverse SERs, double talk, very loud playback and across volume level conditions. We also observe that deep-MVDR residual echo suppression is most likely a better choice over signal processing based MVDR in these complex conditions. We also deduce that multi-stage reconstruction objectives such as playback reconstruction from echo suppressed signals helps our joint training approach by giving further improvements. Our proposed multi-channel approach along with D-MVDR residual echo suppression gave remarkable improvements compared to state-of-the-art approaches. Furthermore, deep-MVDR does perform robust residual echo suppression even when there is moving near-end speaker. It also retained perceptual speech quality of near-end target as per enhanced PESQ scores. Our proposed approach outperforms competitive state-of-the-art approaches best by [ERLE: ([2.1% (abs.)] and [PESQ: ([0.7% (abs.)] for real data TV recordings, and also [ERLE: ([4.7% (abs.)] under adverse SER conditions. This suggests that our proposed approach consistently performed better on both simulated and real-world datasets and is robust to widely observed adverse acoustic ambience conditions discussed in this work.

8. References

[1] C. Breining, P. Dreiscitel, E. Hansler, A. Mader, B. Nitsch, H. Puder, T. Schertler, G. Schmidt, and J. Tilp, "Acoustic Echo Control. An Application Of Very-High-Order Adaptive Filters," *IEEE Signal Processing Magazine*, vol. 16, no. 4, pp. 42–69, 1999.

[2] Y.-S. Choi, H.-C. Shin, and W.-J. Song, "Robust Regularization for Normalized LMS Algorithms," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 53, no. 8, pp. 627–631, 2006.

[3] C. Paleologu, S. Ciochina, J. Benesty, and S. L. Grant, "An Overview On Optimized NLMS Algorithms For Acoustic Echo Cancellation," *EURASIP J. Adv. Signal Processing*, vol. 2015, p. 97, 2015.

[4] S. Ciochină, C. Paleologu, J. Benesty, and S. L. Grant, "An Optimized NLMS Algorithm For Acoustic Echo Cancellation," in *Proc. of International Symposium on Signals, Circuits and Systems (ISSCS)*, Lisbon, Portugal, July 2015, pp. 1–4.

[5] A. Ivry, I. Cohen, and B. Berdugo, "Deep Adaptation Control for Acoustic Echo Cancellation," in *Proc. of ICASSP*, Singapore, May 2022.

[6] C. Lee, J. Shin, and N. Kim, "DNN-Based Residual Echo Suppression," in *Proc. of Interspeech*, Dresden, Germany, Sep 2015, pp. 1775–1779.

[7] J. Franzen and T. Fingscheidt, "Deep Residual Echo Suppression and Noise Reduction: A Multi-Input FCRN Approach in a Hybrid Speech Enhancement System," in *Proc. of ICASSP*, Singapore, May 2022, pp. 666–670.

[8] K. Tan and D. Wang, "A Convolutional Recurrent Neural Network for Real-Time Speech Enhancement," in *Proc. of Interspeech*, Hyderabad, Sep 2018, pp. 3229–3233.

[9] Y. Na, Z. Wang, Z. Liu, B. Tian, and Q. Fu, "Joint Online Multichannel Acoustic Echo Cancellation, Speech Dereverberation and Source Separation," in *Proc. of Interspeech*, Brno, Czech Republic, Sep 2021, pp. 1144–1148.

[10] J. Gu, L. Cheng, X. Sun, J. Li, and Y. Yan, "Residual Echo and Noise Cancellation with Feature Attention Module and Multi-Domain Loss Function," in *Proc. of Interspeech*, Brno, Czech Republic, 2021, pp. 1114–1118.

[11] V. Kothapally, Y. XU, M. Yu, S.-X. ZHANG, and D. Yu, "Joint Neural AEC and Beamforming with Double-Talk Detection," in *Proc. of Interspeech*, Incheon, South Korea, Sep 2022, pp. 2528–2532.

[12] L. Ma, H. Huang, P. Zhao, and T. Su, "Acoustic Echo Cancellation by Combining Adaptive Digital Filter and Recurrent Neural Network," *ArXiv*, vol. abs/2005.09237, 2020.

[13] S. Zhang, Y. Kong, S. Lv, Y. Hu, and L. Xie, "F-T-LSTM Based Complex Network for Joint Acoustic Echo Cancellation and Speech Enhancement," in *Proc. of Interspeech*, 2021, pp. 4758–4762.

[14] *Joint Optimization of Acoustic Echo Cancellation and Adaptive Beamforming*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 19–50. [Online]. Available: https://doi.org/10.1007/3-540-33213-8_2

[15] K. Reindl, Y. Zheng, A. Lombard, A. Schwarz, and W. Kellermann, "An Acoustic Front-end For Interactive TV Incorporating Multichannel Acoustic Echo Cancellation And Blind Signal Extraction," in *Proc. of Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, California, USA, Nov 2010, pp. 1716–1720.

[16] W. Herbordtt, S. Nakamura, and W. Kellermann, "Joint Optimization Of LCMV Beamforming And Acoustic Echo Cancellation For Automatic Speech Recognition," in *Proc. of ICASSP*, vol. 3, Philadelphia, Pennsylvania, USA, 2005, pp. iii/77–iii/80 Vol. 3.

[17] H. Zhang and D. Wang, "A Deep Learning Approach to Multi-Channel and Multi-Microphone Acoustic Echo Cancellation," in *Proc. of Interspeech*, Brno, Czech Republic, 2021, pp. 1139–1143.

[18] ——, "Neural cascade architecture for multi-channel acoustic echo suppression," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2326–2336, 2022.

[19] C. Zhang, J. Liu, H. Li, and X. Zhang, "Neural multi-channel and multi-microphone acoustic echo cancellation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2181–2192, 2023.

[20] Y. Konforti, I. Cohen, and B. Berdugo, "Multichannel acoustic echo cancellation with beamforming in dynamic environments," *IEEE Open Journal of Signal Processing*, vol. 4, pp. 479–488, 2023.

[21] B. Kwon, Y. Park, and Y.-s. Park, "Analysis Of The GCC-PHAT Technique For Multiple Sources," pp. 2070–2073, Oct 2010.

[22] C. K. Reddy, E. Beyrami, J. Pool, R. Cutler, S. Srinivasan, and J. Gehrke, "A Scalable Noisy Speech Dataset and Online Subjective Test Framework," *Proc. of Interspeech*, pp. 1816–1820, 2019.

[23] R. Cutler, A. Saabas, T. Parnamaa, M. Purin, H. Gamper, S. Braun, K. Sorensen, and R. Aichner, "ICASSP 2022 Acoustic Echo Cancellation Challenge," in *Proc. of ICASSP*, Singapore, May 2022.

[24] R. Scheibler, E. Bezzam, and I. Dokmanic, "Pyroomacoustics: A Python Package For Audio Room Simulations And Array Processing Algorithms," *CoRR*, vol. abs/1710.04196, 2017. [Online]. Available: http://arxiv.org/abs/1710.04196

[25] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR Corpus Based On Public Domain Audio Books," in *Proc. of ICASSP*, Brisbane, Queensland, Australia, april 2015, pp. 5206–5210.

[26] C. Paleologu, S. Ciochina, and J. Benesty, "Variable Step-Size NLMS Algorithm for Under-Modeling Acoustic Echo Cancellation," *IEEE Signal Processing Letters*, vol. 15, pp. 5–8, 2008.