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

A Deep Learning Approach to Multi-Channel and Multi-Microphone Acoustic Echo Cancellation

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
{zhang.6720, wang.77}@osu.edu

Abstract

Building on deep learning based acoustic echo cancellation (AEC) in the single-loudspeaker (single-channel) and single-microphone setup, this paper investigates multi-channel (multi-loudspeaker) AEC (MCAEC) and multi-microphone AEC (MMAEC). A convolutional recurrent network (CRN) is trained to predict the near-end speech from microphone signals with far-end signals used as additional information. We find that the deep learning based MCAEC approach avoids the non-uniqueness problem in traditional MCAEC algorithms. For the AEC setup with multiple microphones, rather than employing AEC for each microphone, we propose to train a single network to achieve echo removal for all microphones. Combining deep learning based AEC with supervised beamforming further improves the system performance. Experimental results show the effectiveness of deep learning approach to MCAEC and MMAEC. Furthermore, deep learning based methods are capable of removing echo and noise simultaneously and work well in the presence of nonlinear distortions.

Index Terms: acoustic echo cancellation, deep learning, multi-channel AEC, multi-microphone AEC, nonlinearity

1. Introduction

Acoustic echo cancellation (AEC) is the task of removing undesired echoes that result from the coupling between a loudspeaker and a microphone in a communication system [1]. Modern hands-free communication devices are usually equipped with multiple microphones and loudspeakers. The availability of additional devices also elevates the need for enhanced sound quality and realism, which can hardly be satisfied with single-channel AEC. Therefore, it is necessary to design AEC for multiple loudspeakers and/or microphones, which leads to the study of MCAEC and MMAEC. MCAEC and MMAEC present additional challenges and opportunities compared to single-channel AEC and have received considerable attention recently.

Multi-channel AEC refers to the setup with at least two loudspeakers or channels (stereophonic sound). Although conceptually similar, MCAEC is fundamentally different from single-channel AEC and a straightforward generalization of single-channel AEC does not result in satisfactory performance because of the non-uniqueness problem [2]. This problem is due to the correlation between loudspeaker signals. As a result, the convergence of adaptive technique could be degraded and the echo paths cannot be determined uniquely [2]. Many methods have been proposed to circumvent this problem [3, 4, 5, 6], among which coherence reduction methods are most commonly used. Such methods, however, inevitably degrade sound quality, and a compromise must be made between enhanced convergence and sound quality corruption [2, 7].

MMAEC is required for situations in which multiple microphones are present and beamforming techniques are usually combined with AEC for efficient reduction of noise and acoustic echoes. The most straightforward ways of combining these two processing modules are applying AEC separately for each microphone signal before beamforming or applying a single-microphone AEC to the output of a beamformer [8]. In general, the former scheme outperforms the latter one [9, 10]. Other algorithms employ relative echo transfer functions [11, 12] or joint optimization strategies [13, 14] to improve the MMAEC performance. However, efficient combinations of AEC and beamforming are still challenging and many of the strategies exhibit convergence deficiencies [7].

Recently, deep learning based methods have been proposed for solving AEC problems and have shown to be effective for echo and noise removal, especially in situations with nonlinear distortions [15, 16, 17, 18]. On the basis of the deep learning based single-channel AEC approach, we investigate AEC setups with multiple loudspeakers and microphones. The CRN [19] based method is introduced to address MCAEC and MMAEC problems. Evaluation results show that the proposed method effectively remove acoustic echo and background noise in the presence of nonlinear distortions.

The proposed work has four major advantages over traditional methods. First, instead of estimating echo paths, deep learning based MCAEC works by directly estimating near-end speech, which intrinsically avoids the non-uniqueness problem. Second, although there are multiple acoustic paths in the MCAEC and MMAEC setups, the deep learning based approach can naturally address the problem with model training, rather than employing a separate AEC module for each echo path. Third, combining deep learning based AEC and deep learning based beamforming elevates AEC performance remarkably. Fourth, deep learning based methods can remove echo and noise simultaneously in the presence of nonlinear distortions.

The remainder of this paper is organized as follows. Section 2 presents the deep learning approach to MCAEC and MMAEC. Experiments and evaluation results are given in Section 3. Section 4 concludes the paper.

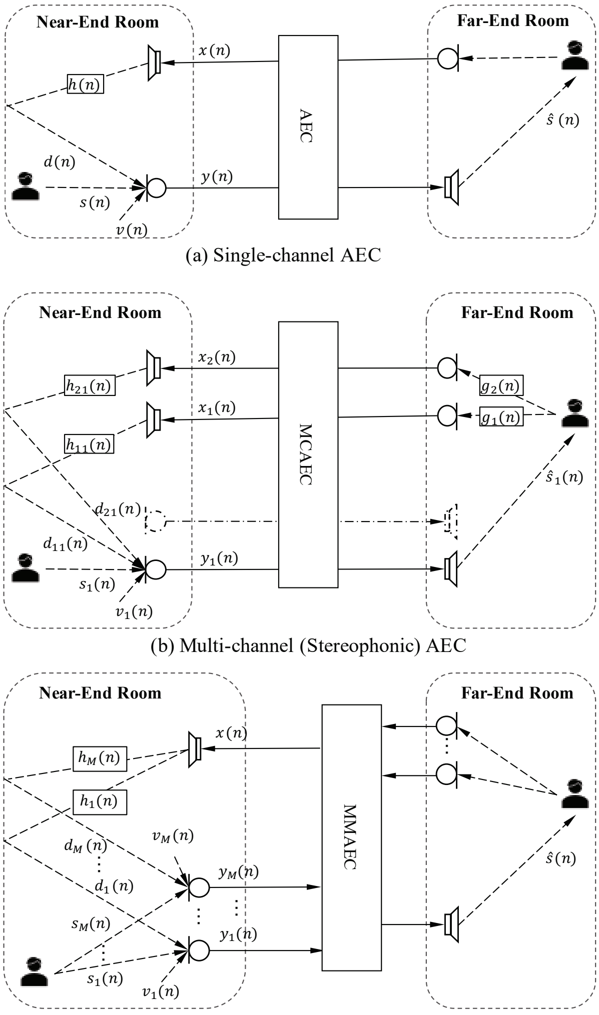
2. Method Description

2.1. Deep learning based AEC

As is shown in Fig. 1(a), the microphone signal $y(n)$ in the single-channel AEC setup is a mixture of echo $d(n)$, near-end speech $s(n)$, and background noise $v(n)$:

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

http://dx.doi.org/10.21437/Interspeech.2021-1508



(a) Single-channel AEC

(b) Multi-channel (Stereophonic) AEC

(c) Multi-microphone AEC

Figure 1: *Diagrams of conventional (1) single-channel AEC setup, (2) Multi-channel (Stereophonic) AEC setup, and (c) Multi-microphone AEC setup.*

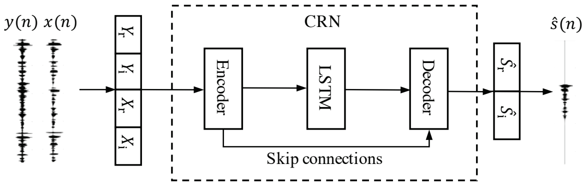


Figure 2: *CRN based AEC method. Subscripts r and i denote real and imaginary spectra of signals, respectively.*

where n indexes a time sample, and echo is generated by convolving a loudspeaker signal with a room impulse response (RIR) ($h(n)$). The echo $d(n)$ is typically a linear or nonlinear transform of the far-end signal $x(n)$. We formulate AEC as a supervised speech separation problem and the overall approach is to estimate the near-end speech from microphone signal with far-end signal used as additional information. The diagrams of deep learning based methods are shown in Fig. 2. The input signals, sampled at 16 kHz, are windowed into 20 ms frames with a 10-ms overlap between consecutive frames. Then a 320-point short time Fourier transform (STFT) is applied to each frame to extract the real and imaginary spectra of signals, which are denoted as $*_r$ and $*_i$, respectively.

A CRN is trained for complex spectral mapping. As is shown in Fig. 2, it estimates the real and imaginary spectrograms of near-end speech ($\hat{S}_r$ and $\hat{S}_i$) from the real and imaginary spectrograms of microphone signal and far-end signal (Y_r, Y_i, X_r, and X_i). Then $\hat{S}_r$ and $\hat{S}_i$ are sent to the inverse short time Fourier transform to derive an estimated near-end signal $\hat{s}(n)$. Hence, it is capable of enhancing both magnitude and phase responses simultaneously and $\hat{s}(n)$ resynthesized achieves better speech quality. The CRN is an encoder-decoder architecture. Specifically, the encoder and decoder comprise five convolutional layers and five deconvolutional layers, respectively. Between them is a two-layer LSTM with a group strategy, where the group number is set to 2. A detailed description the CRN architecture is provided in [19].

2.2. Deep learning for MCAEC

Without loss of generality, let us take stereophonic AEC as an example to study deep learning based MCAEC. The signal model is given in Fig. 1(b) where the stereophonic signals, $x_1(n)$ and $x_2(n)$ are transmitted to loudspeakers and then coupled to one of the microphones, $h_{ij}(n)$ denotes the echo path from loudspeaker i to microphone j. The signal picked up by microphone j is composed of two echo signals $d_{1j}(n)$, $d_{2j}(n)$, near-end speech $s_j(n)$, and background noise $v_j(n)$:

$$y_j(n) = \textstyle\sum_{i=1}^{2} d_{ij}(n) + s_j(n) + v_j(n), \quad j = 1, 2 \qquad (2)$$

Deep learning based MCAEC works by estimating the target $s_j(n)$ given $y_j(n)$, $x_1(n)$, and $x_2(n)$ as inputs. Specifically, we use $[Y_{jr}, Y_{ji}, X_{1r}, X_{1i}, X_{2r}, X_{2i}]$ as inputs and train the network to directly estimate $[S_{jr}, S_{ji}]$ from them. And there is not need to do any de-correlation preprocessing to the stereophonic signals. Therefore, deep learning based MCAEC method can avoid the non-uniqueness problem, which traditional adaptive filter-based methods often suffer from. In the proposed method, the training signals are generated by randomly selecting j from $\{1, 2\}$, i.e. the model is exposed to signals picked up by the two microphones in MCAEC during training. A model trained this way is able to achieve echo removal for both microphones in the system.

2.3. Deep learning for MMAEC

Considering an MMAEC setup with one loudspeaker and M microphones, as is shown in Fig. 1(b). The signal picked up by microphone j is

$$y_j(n) = d_j(n) + s_j(n) + v_j(n), \quad j = 1, 2, \cdots M \qquad (3)$$

Different strategies for solving MMAEC problems have been discussed in [8, 9, 10]. In this paper, we focus on the most straightforward one, which is to apply AEC separately for each microphone signal before beamforming [8]. Therefore, this MMAEC strategy does not structurally differ from the single-microphone case in terms of AEC. Based on this strategy, traditional MMAEC methods need to update M AEC modules separately to achieve echo removal for all the M microphones in the array. While deep learning based MMAEC can be trained to achieve this with a single network rather than training separate networks for each microphone. During training, we use $y_j(n)$ and $x(n)$ as inputs and set the corresponding near-end

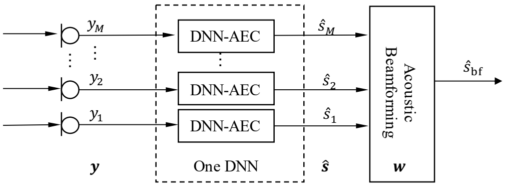


Figure 3: *Diagram of combining deep learning based AEC with deep learning based beamforming for further enhancement.*

speech $s_j(n)$ as the training target. The training signals are generated by randomly choosing j from $\{1, 2, \cdots, M\}$. A model trained this way is able to achieve echo removal for all the microphones in the array.

Once the model is trained, the outputs of the model can be used for deep learning based minimum variance distortionless response (MVDR) beamforming [20]. Choosing the first microphone in the array as reference microphone, the MVDR beamformer can be constructed as:

$$\hat{\boldsymbol{w}}(f) = \frac{\hat{\Phi}_N^{-1}(f)\hat{\boldsymbol{c}}(f)}{\hat{\boldsymbol{c}}(f)^H\hat{\Phi}_N^{-1}(f)\hat{\boldsymbol{c}}(f)} \tag{4}$$

where $(\cdot)^H$ denotes conjugate transpose, $\hat{\Phi}_N(f)$ is the estimated covariance matrix of overall interference (acoustic echo and background noise), $\hat{\boldsymbol{c}}(f)$ is the estimated steering vector, which is estimated as the principal eigenvector of the estimated speech covariance matrix $\hat{\Phi}_s(f)$ [20, 21]. The estimated covariance matrices of speech and overall interference are obtained from the output of deep learning based MMAEC as

$$\hat{\Phi}_S(f) = \tfrac{1}{T}\textstyle\sum_t \hat{\boldsymbol{S}}(t,f)\hat{\boldsymbol{S}}^H(t,f) \tag{5}$$

$$\hat{\Phi}_N(f) = \tfrac{1}{T}\textstyle\sum_t \hat{\boldsymbol{N}}(t,f)\hat{\boldsymbol{N}}^H(t,f) \tag{6}$$

where $\hat{\boldsymbol{S}}(t,f)$ is the STFT representation of estimated speech signals and $\hat{\boldsymbol{N}}(t,f)$ is the estimated overall interference obtained as $\boldsymbol{Y}(t,f) - \hat{\boldsymbol{S}}(t,f)$, T is the total number of frames used in the summation.

The beamformer is usually applied on microphone signal $\boldsymbol{Y}(t,f)$ and the enhancement results are calculated from

$$Y_{\text{bf}}(t,f) = \hat{\boldsymbol{w}}^H(f)\boldsymbol{Y}(t,f) \tag{7}$$

Considering that MVDR beamformer performs spatial filtering to maintain signals from the desired direction while suppressing interferences from other directions, we proposed to use it as a post-filter for further enhancement. The overall structure of the deep learning based MMAEC is shown in Fig. 3. It is implemented by feeding the output of deep learning based AEC forward to the deep learning based beamformer with the latter calculated using the same network. The further enhanced output is obtained using

$$\hat{S}_{\text{bf}}(t,f) = \hat{\boldsymbol{w}}^H(f)\hat{\boldsymbol{S}}(t,f) \tag{8}$$

3. Experiments

3.1. Experiment setting

The simulation setups for evaluation are designed as follows. The near-end and far-end speech signals are generated using the TIMIT dataset [22] by following the same way provided in [17]. RIRs are generated using the image method [23]. To investigate RIRs generalization, we simulate 20 different rooms of size $a \times b \times c$ m (width×length×height) for training mixtures, where $a = [4, 6, 8, 10], b = [5, 7, 9, 11, 13], c = 3$. For MCAEC setup, the two microphones and the two loudspeakers are positioned at $(a/2, b/2 + 0.05, c/2)$ m, $(a/2, b/2 - 0.05, c/2)$ m, $(a/2, b/2 + 0.6, c/2 + 0.5)$ m, and $(a/2, b/2 - 0.6, c/2 + 0.5)$ m, respectively. The near-end speaker is put at 20 random positions in each room with 1 meter apart from the center of the microphones. The setup of MMAEC consists of a uniform linear array with four microphones and one loudspeaker. The center of the microphone array is positioned at the center of the room with 4 cm inter-microphone distance. Twenty pairs of positions are simulated randomly for the loudspeaker and the near-end speaker in each room, and the distance from the loudspeaker and the near-end speaker to the center of the array are set to 0.6 m and 1 m, respectively. The reverberation time (T_{60}) is randomly selected from $\{0.2, 0.3, 0.4, 0.5, 0.6\}$ s. For testing, we simulate three untrained rooms of size $3 \times 4 \times 3$ m (Room 1), $5 \times 6 \times 3$ m (Room 2), $11 \times 14 \times 3$ m (Room3), and set T_{60} to 0.35 s to generate test RIRs for both MCAEC and MMAEC setups.

The most common nonlinear distortion generated by a loudspeaker is the saturation type nonlinearity, which is usually simulated using the scaled error function (SEF) [24, 25]:

$$f_{\text{SEF}}(x) = \int_0^x e^{-\frac{z^2}{2\eta^2}}dz \tag{9}$$

where x is the input to the loudspeaker, η^2 represents the strength of nonlinearity. The SEF becomes linear as η^2 tends to infinity and becomes a hard limiter as η^2 tends to zero. To investigate the robustness of the proposed method against nonlinear distortions, four loudspeaker functions are used during the training stage: $\eta^2 = 0.1$ (severe nonlinearity), $\eta^2 = 1$ (moderate nonlinearity), $\eta^2 = 10$ (soft nonlinearity), and $\eta^2 = \infty$ (linear).

Babble noise from NOISEX-92 dataset [26] is used as the background noise and the algorithm proposed in [27] is employed to make the noise diffuse. The diffuse babble noise is then split into two parts, the first 80% of it is used for training and the remaining is used for testing.

We create 20000 training mixtures and 100 test mixtures for both MCAEC and MMAEC setups. A loudspeaker signal is generated using a randomly selected far-end signal and a loudspeaker function. Then a loudspeaker signal is convolved with a randomly chosen training RIR for loudspeaker to generate an echo. A randomly chosen near-end utterance is convolved with an RIR for near-end speaker and then mixed with the echo at a signal-to-echo ratio (SER) randomly chosen from $\{-6, -3, 0, 3, 6\}$ dB. The diffuse babble noise is added to the mixture at a signal-to-noise ratio (SNR) randomly chosen from $\{8, 10, 12, 14\}$ dB. The SER and SNR, which are evaluated during double-talk periods, are defined as:

$$\text{SER} = 10\log_{10}\left[\textstyle\sum_n s^2(n)/\sum_n d^2(n)\right] \tag{10}$$

$$\text{SNR} = 10\log_{10}\left[\textstyle\sum_n s^2(n)/\sum_n v^2(n)\right] \tag{11}$$

Test mixtures are created similarly but using different utterances, noises, RIRs, SERs and SNRs.

The AMSGrad optimizer [28] and the mean squared error (MSE) cost function are used to train CRN. The network is trained for 30 epochs with a learning rate of 0.001. The performance of MCAEC and MMAEC is evaluated in terms of

Table 1: *Performance of MCAEC methods in the presence of double-talk, background noise with 3.5 dB SER, 10 dB SNR, $\eta^2 = \infty$ (linear system).*

	ERLE			PESQ		
RIRs	Room1	Room2	Room3	Room1	Room2	Room3
Unprocessed	-	-	-	2.12	2.13	2.17
SJONLMS	7.54	7.63	7.62	2.41	2.45	2.47
SJONLMS-PF	19.04	18.88	18.67	2.45	2.48	2.53
CRN	**32.68**	**35.34**	**41.31**	**2.62**	**2.83**	**2.98**

Table 2: *Performance of MMAEC methods in the presence of double-talk, background noise with 3.5 dB SER, 10 dB SNR, $\eta^2 = \infty$ (linear system).*

		ERLE			PESQ		
RIRs		Room1	Room2	Room3	Room1	Room2	Room3
Unprocessed		-	-	-	2.04	2.09	2.10
JONLMS		6.94	6.95	6.93	2.43	2.45	2.48
JONLMS-IBF		17.61	16.76	15.52	2.70	2.63	2.66
CRN	$\hat{s}$	25.92	32.94	33.99	2.66	2.89	**2.94**
	y_{bf}	2.77	5.48	2.24	2.18	2.41	2.21
	$\hat{s}_{\text{bf}}$	**27.57**	**36.92**	**34.11**	**2.75**	**2.98**	2.89

utterance-level echo return loss enhancement (ERLE) [1] for single-talk periods and perceptual evaluation of speech quality (PESQ) [29] for double-talk periods.

3.2. Performance of MCAEC methods

We first evaluate the performance of deep learning based MCAEC. The proposed methods are compared with the stereophonic version of joint-optimized normalized least mean square algorithm [30] equipped with a coherence reduction technique proposed in [31] (SJONLMS). And post-filtering (PF) [32] is employed to further suppress noises and residual echo (SJONLMS-PF). The parameters of SJONLMS and PF are set accordingly to the values given in [30, 31, 32]. The comparison results are given in Table 1. In general, the proposed CRN based MCAEC method outperforms conventional methods and the performance generalizes well to untrained RIRs.

3.3. Performance of MMAEC methods

This part studies the performance of deep learning based MMAEC. We employ single-channel JONLMS [30] for each microphone in the array as a baseline and then combine the outputs with the ideal MVDR beamformer (JONLMS-IBF) for solving the MMAEC problem. The ideal MVDR beamformer (IBF) is calculated by substituting the true speech and interference components of the microphone signal ($\boldsymbol{S}(t, f)$ and $\boldsymbol{N}(t, f)$) into (6), (7), and (5). Therefore, it can be regarded as a stronger baseline compared to other MVDR beamformers. Three results are provided for each deep learning based method, in which $\hat{s}$ is the output of the reference microphone, y_{bf} and $\hat{s}_{\text{bf}}$ are, respectively, the time-domain beamformed microphone signal and beamformed enhanced signal introduced in Section 2.3. The comparison results are given in Table 2. As can be seen from the table, deep learning based methods outperform traditional MMAEC methods in terms of ERLE and PESQ. Single-channel outputs of deep learning based methods ($\hat{s}$) are good enough for echo and noise removal while combining deep learning based beamformer as a post-filter ($\hat{s}_{\text{bf}}$) further improves the overall performance in most of the cases.

Table 3: *Performance of the proposed method in the presence of double-talk, background noise and nonlinear distortions with 3.5 dB SER, 10 dB SNR, Room2, $\eta^2 = 0.1$, and $\eta^2 = 0.5$.*

			ERLE		PESQ	
Nonlinearity			$\eta^2 = 0.1$	$\eta^2 = 0.5$	$\eta^2 = 0.1$	$\eta^2 = 0.5$
MCAEC	Unprocessed		-	-	2.11	2.13
	CRN		34.86	34.72	2.82	2.83
MMAEC	Unprocessed		-	-	2.08	2.08
	CRN	$\hat{s}$	33.15	33.05	2.89	2.88
		y_{bf}	5.49	5.46	2.40	2.40
		$\hat{s}_{\text{bf}}$	36.84	36.83	2.99	2.99

Table 4: *Performance of the proposed method under untrained speakers and moveable speakers conditions with 3.5 dB SER,10 dB SNR, Room 2, and $\eta^2 = 0.1$.*

			Untrained speakers		Moveable speakers	
			ERLE	PESQ	ERLE	PESQ
MCAEC	Unprocessed		-	2.10	-	2.10
	CRN		35.57	2.83	35.65	2.76
MMAEC	Unprocessed		-	2.06	-	2.06
	CRN	$\hat{s}$	33.53	2.89	33.48	2.80
		y_{bf}	5.52	2.37	3.46	2.34
		$\hat{s}_{\text{bf}}$	37.10	2.99	34.97	2.93

3.4. Robustness of the proposed methods

This part tests the robustness of deep learning based methods to nonlinear distortions, untrained speakers, and moveable speakers. The results of CRN based MCAEC and MMAEC in situations with the saturation type nonlinear distortions are given in Table 3. It is seen that deep learning based methods can be trained to handle both linear and nonlinear cases and the performance generalizes well to untrained nonlinearity ($\eta^2 = 0.5$). Table 4 shows the behavior of the proposed method when tested with untrained speakers and moveable speakers. The test signals of untrained speakers are created by randomly selecting 10 pairs of untrained speakers from the TIMIT dataset. As for moveable speakers, we simulate the case by changing the position of the near-end speaker (for example, from $(1.5, 3, 1.5)$ m to $(1.7, 2.8, 2.0)$ m) at the middle point of a near-end utterance and using the corresponding RIRs to generate a near-end speech. The results in this table demonstrate high robustness of the proposed methods.

4. Conclusion

We have proposed a deep learning approach to MCAEC and MMAEC. Our approach overcomes the limitations of traditional methods and produces remarkable performance in terms of ERLE and PESQ. Evaluation results show the effectiveness of CRN based methods for removing echo and noise in cases with and without nonlinear distortions, and the performance generalizes well to untrained RIRs. Moreover, the proposed methods can be extended to handle a general AEC setup with an arbitrary number of microphones and an arbitrary number of loudspeakers, which will be demonstrated in future research.

5. Acknowledgements

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

6. References

[1] G. Enzner, H. Buchner, A. Favrot, and F. Kuech, "Acoustic echo control," in *Academic press library in signal processing: image, video processing and analysis, hardware, audio, acoustic and speech Processing*. Academic Press, 2014.

[2] M. M. Sondhi, D. R. Morgan, and J. L. Hall, "Stereophonic acoustic echo cancellation-an overview of the fundamental problem," *IEEE Signal processing letters*, vol. 2, no. 8, pp. 148–151, 1995.

[3] J. Benesty, F. Amand, A. Gilloire, and Y. Grenier, "Adaptive filtering algorithms for stereophonic acoustic echo cancellation," in *1995 ICASSP*, vol. 5. IEEE, 1995, pp. 3099–3102.

[4] S. Shimauchi, S. Makino, and J. Kojima, "Method and apparatus for multi-channel acoustic echo cancellation," Aug. 26 1997, uS Patent 5,661,813.

[5] M. Schneider and W. Kellermann, "Multichannel acoustic echo cancellation in the wave domain with increased robustness to nonuniqueness," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 518–529, 2016.

[6] J. Franzen and T. Fingscheidt, "An efficient residual echo suppression for multi-channel acoustic echo cancellation based on the frequency-domain adaptive Kalman filter," in *2018 ICASSP*. IEEE, 2018, pp. 226–230.

[7] M. Luis Valero, "Acoustic echo reduction for multiple loudspeakers and microphones: Complexity reduction and convergence enhancement," Ph.D. dissertation, Friedrich-Alexander-University of Erlangen-Nürnberg, 2019.

[8] W. Kellermann, "Strategies for combining acoustic echo cancellation and adaptive beamforming microphone arrays," in *1997 ICASSP*, vol. 1. IEEE, 1997, pp. 219–222.

[9] W. Herbordt and W. Kellermann, "Limits for generalized sidelobe cancellers with embedded acoustic echo cancellation," in *2001 ICASSP*, vol. 5. IEEE, 2001, pp. 3241–3244.

[10] S. Doclo, M. Moonen, and E. De Clippel, "Combined acoustic echo and noise reduction using GSVD-based optimal filtering," in *2000 ICASSP*, vol. 2. IEEE, 2000, pp. II1061–II1064.

[11] G. Reuven, S. Gannot, and I. Cohen, "Joint noise reduction and acoustic echo cancellation using the transfer-function generalized sidelobe canceller," *Speech communication*, vol. 49, no. 7-8, pp. 623–635, 2007.

[12] M. L. Valero and E. A. Habets, "Multi-microphone acoustic echo cancellation using relative echo transfer functions," in *2017 WASPAA*. IEEE, 2017, pp. 229–233.

[13] W. Herbordt, W. Kellermann, and S. Nakamura, "Joint optimization of LCMV beamforming and acoustic echo cancellation," in *2004 12th European Signal Processing Conference*. IEEE, 2004, pp. 2003–2006.

[14] W. Herbordt and W. Kellermann, "Gsaecacoustic echo cancellation embedded into the generalized sidelobe canceller," in *2000 10th European Signal Processing Conference*. IEEE, 2000, pp. 1–4.

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

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

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

[18] 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.

[19] K. Tan and D. L. Wang, "A convolutional recurrent neural network for real-time speech enhancement," in *Interspeech*, 2018, pp. 3229–3233.

[20] J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, "BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 444–451.

[21] X. Zhang, Z. Wang, and D. L. Wang, "A speech enhancement algorithm by iterating single- and multi-microphone processing and its application to robust ASR," in *2017 ICASSP*. IEEE, 2017, pp. 276–280.

[22] 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.

[23] 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.

[24] F. Agerkvist, "Modelling loudspeaker non-linearities," in *Audio Engineering Society Conference: 32nd International Conference: DSP For Loudspeakers*. Audio Engineering Society, 2007.

[25] H. Zhang and D. L. Wang, "A deep learning approach to active noise control," in *2020 INTERSPEECH in press*, 2020.

[26] 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.

[27] E. A. Habets, I. Cohen, and S. Gannot, "Generating nonstationary multisensor signals under a spatial coherence constraint," *The Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 2911–2917, 2008.

[28] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of adam and beyond," *arXiv preprint arXiv:1904.09237*, 2019.

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

[30] 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, p. 97, 2015.

[31] M. Djendi, "An efficient stabilized fast Newton adaptive filtering algorithm for stereophonic acoustic echo cancellation SAEC," *Computers & Electrical Engineering*, vol. 38, no. 4, pp. 938–952, 2012.

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