ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9414623

WEIGHTED RECURSIVE LEAST SQUARE FILTER AND NEURAL NETWORK BASED RESIDUAL ECHO SUPPRESSION FOR THE AEC-CHALLENGE

Ziteng Wang[†‡], *Yueyue Na*[†], *Zhang Liu*[†], *Biao Tian*[†], *Qiang Fu*[†‡]

[†]Machine Intelligence Technology, Alibaba Group
[‡]Beijing Sound Connect Technology

ABSTRACT

This paper presents a real-time Acoustic Echo Cancellation (AEC) algorithm submitted to the AEC-Challenge. The algorithm consists of three modules: Generalized Cross-Correlation with PHAse Transform (GCC-PHAT) based time delay compensation, weighted Recursive Least Square (wRLS) based linear adaptive filtering and neural network based residual echo suppression. The wRLS filter is derived from a novel semi-blind source separation perspective. The neural network model predicts a Phase-Sensitive Mask (PSM) based on the aligned reference and the linear filter output. The algorithm achieved a mean subjective score of 4.00 and ranked 2nd in the AEC-Challenge.

Index Terms— AEC-Challenge, weighted RLS, residual echo suppression, deep neural network

1. INTRODUCTION

Acoustic Echo Cancellation (AEC) plays an essential part in full-duplex speech communication systems. The goal of AEC is no echo leakage when there is loudspeaker signal (far end) and no speech distortion when the users talk (near end). It has been a challenging problem since the earlier days of telecommunication [1]. A practical acoustic echo cancellation solution, e.g. the one in the WebRTC project [2], usually consists of three modules: Time Delay Compensation (TDC), linear adaptive filtering and Non-Linear Processing (NLP).

Time delay compensation is necessary, especially in real systems where microphone signal capturing and loudspeaker signal rendering are handled by different threads and the sample clocks may not be synchronized. Typical delays between the far end and near end signals range from 10 ms to 500 ms. Though in theory, the linear adaptive filter can handle any delay by having a sufficient number of filter taps. TDC could benefit the performance by avoiding over-parameterization and speeding up convergence. Time delay estimation methods include the Generalized Cross-Correlation with PHAse Transform (GCC-PHAT) algorithm [3] and audio fingerprinting technology [4].

Linear adaptive filters, such as Normalized Least Mean Square (NLMS) filters [5] and Kalman filters [6], can be designed either in the time domain or in the frequency domain. For the best performance possible, the filter length should be long enough to cover the whole echo path, which could be thousands of taps in the time domain. Frequency Domain Adaptive Filter (FDAF) [7] are more often chosen for computational savings and better modeling statistics.

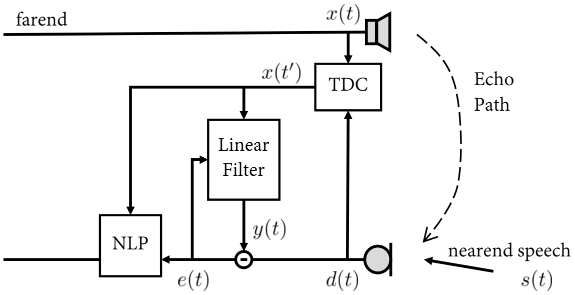

Fig. 1. A typical acoustic echo cancellation solution.

NLP is introduced as a complement to linear filtering to suppress residual echos. The methods are generally adapted from noise reduction techniques, e.g. the multi-frame Wiener filter [8]. Many recent studies also adopt deep learning methods for residual echo suppression [9, 10, 11, 12] and report reasonable objective scores on synthetic datasets. One concern is that the neural network models may degrade significantly in real applications. The AEC-Challenge [13] is thus organized to stimulate research in this area by providing recordings from more than 2,500 real audio devices and human speakers in real environments. The evaluation is based on the average P.808 Mean Opinion Score (MOS) [14] achieved across all different single talk and double talk scenarios.

This paper describes our submission to the AEC-Challenge, which consists of three cascading modules: GCC-PHAT for time delay compensation, weighted Recursive Least Square (wRLS) for linear filtering, and a Deep Feedforward Sequential Memory Network (Deep-FSMN) [15] for residual echo suppression. The wRLS filter is derived from a novel semi-blind source separation perspective and is shown to be double talk friendly. The algorithm proved its efficacy in the

Challenge and it is described in the following section.

2. THE PROPOSED ALGORITHM

As in Figure 1, the captured signal at time t is expressed as:

$$d(t) = x(t) * a(t) + s(t) + v(t) \tag{1}$$

where $x(t), s(t)$ and $v(t)$ are respectively the the far end signal, the near end speech signal and the signal modeling error. $a(t)$ denotes the echo path and $*$ denotes convolution. It is assumed $v(t) = 0$ in the following for simplicity. The frequency representations of d, x, a, s are respectively denoted as D, X, A, S.

2.1. Time Delay Compensation

The GCC-PHAT algorithm is applied first to align the far end and near end signals. The generalized cross correlation is defined as $\Phi_{t,f} = E[X_{t,f}D^*_{t,f}]$ with $E[\cdot]$ denoting expectation, f the frequency index and $(\cdot)^*$ the conjugate of a variable. The online implementation is given by:

$$\Phi_{t,f} = \alpha\Phi_{t-1,f} + (1-\alpha)X_{t,f}D^*_{t,f} \tag{2}$$

where α is a smoothing parameter. The relative delay τ is obtained by performing Inverse Fast Fourier Transform (IFFT) and finding the index of the maximum:

$$\tau = \underset{\tau}{\text{argmax}}\ \text{IFFT}\left(\frac{\Phi_{t,f}}{|\Phi_{t,f}|}\right) \tag{3}$$

2.2. wRLS Filtering

Linear filtering is performed in the frequency domain on the time-aligned signals $x(t')$ and $d(t)$. Suppose an echo path of L taps, the signal model is reformulated as:

$$\begin{bmatrix} D_{t,f} \\ \mathbf{x}_{L,f} \end{bmatrix} = \begin{bmatrix} 1 & \mathbf{a}^H_{L,f} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} S_{t,f} \\ \mathbf{x}_{L,f} \end{bmatrix} \tag{4}$$

where $\mathbf{x}_{L,f} = [X(t', f), X(t'-1, f), ..., X(t'-L+1, f)]^T$ and $\mathbf{a}_{L,f} = [A(t, f), A(t-1, f), ..., A(t-L+1, f)]^T$ with $(\cdot)^T$ denoting transpose and $(\cdot)^H$ Hermitian transpose. $\mathbf{I}$ is a unitary matrix of order L. The near end speech can be separated by:

$$\begin{bmatrix} \hat{S}_{t,f} \\ \mathbf{x}_{L,f} \end{bmatrix} = \mathbf{B}_f \begin{bmatrix} D_{t,f} \\ \mathbf{x}_{L,f} \end{bmatrix} \tag{5}$$

where $\hat{(\cdot)}$ denotes the estimate of a variable and $\mathbf{B}_f$ is termed the unmixing matrix.

Equation (5) clearly defines a semi-blind source separation problem. Assuming independence of $\{D_{t,f}, \mathbf{x}_{L,f}\}$, the unmixing matrix has this unique form as:

$$\mathbf{B}_f = \begin{bmatrix} 1 & \mathbf{w}^H_{L,f} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \tag{6}$$

which can be solved by the well established source source separation algorithms, such as the Independent Component Analysis (ICA) and auxiliary-function based (Aux-)ICA algorithms [16]. The Aux-ICA solution is briefly described as follows and a detailed derivation can be found in [17].

The Kullback-Leibler divergence is introduced as the independence measure

$$J(\mathbf{B}_f) = \int_{S_{t,f}} \int_{\mathbf{x}_{L,f}} p(S_{t,f}, \mathbf{x}_{L,f}) \log \frac{p(S_{t,f}, \mathbf{x}_{L,f})}{q(S_{t,f}, \mathbf{x}_{L,f})} \tag{7}$$

where $p(\cdot)$ represents the source Probability Density Function (PDF) and $q(\cdot)$ the product of approximated PDF of individual sources. The loss is upper bounded by the auxiliary loss function

$$Q(\mathbf{B}_f, \mathbf{C}_f) = \sum_{i=1}^{L+1} \mathbf{b}^H_{i,f}\mathbf{C}_{i,f}\mathbf{b}_{i,f} + const. \tag{8}$$

where $\mathbf{b}^H_{i,f}$ is the i-th row vector of $\mathbf{B}_f$ and the auxiliary variable

$$\mathbf{C}_{i,f} = E[\frac{G'(r_{i,t,f})}{r_{i,t,f}}\mathbf{x}_{t,f}\mathbf{x}^H_{t,f}] \tag{9}$$

with $\mathbf{x}_{t,f} = [D_{t,f}, \mathbf{x}^T_{L,f}]^T$ and $r_{i,t,f}$ the i-th separated source. $G(r)$ is called the contrast function and has a relationship $G(r) = -\log p(r)$.

Equation (8) can be minimized in terms of $\mathbf{b}_{1,f}$ as:

$$\begin{aligned} \mathbf{b}_{1,f} &= [\mathbf{B}_f\mathbf{C}_{1,f}]^{-1}\mathbf{i}_1 \\ &= \mathbf{C}^{-1}_{1,f}\mathbf{i}_1. \end{aligned} \tag{10}$$

with $\mathbf{i}_1 = [1, 0, ..., 0]^T$ a $L+1$ dimensional vector. Further by applying block matrix inversion of $\mathbf{C}_{1,f}$, the unmixing filter coefficients are given by

$$\mathbf{w}_{L,f} = -\mathbf{R}^{-1}_{L,f}\mathbf{r}_{L,f} \tag{11}$$

where

$$\begin{aligned} \mathbf{R}_{L,f} &= E[\frac{G'(r)}{r}\mathbf{x}_{L,f}\mathbf{x}^H_{L,f}], \\ \mathbf{r}_{L,f} &= E[\frac{G'(r)}{r}\mathbf{x}_{L,f}D^*_{t,f}]. \end{aligned} \tag{12}$$

The separated near end speech is obtained as:

$$\hat{S}_{t,f} = D_{t,f} + \mathbf{w}^H_f\mathbf{x}_{L,f}. \tag{13}$$

Equation (11) stands for a weighted RLS filter, in which the correlation weighting factor is determined by the underlying near end source PDF. In literature, a general super-Gaussian source PDF has the form of

$$G(D_{t,f}) = (\frac{D_{t,f}}{\eta})^\beta, \quad 0 < \beta \leq 2 \tag{14}$$

where a shape parameter of $\beta \in [0.2, 0.4]$ is suggested.

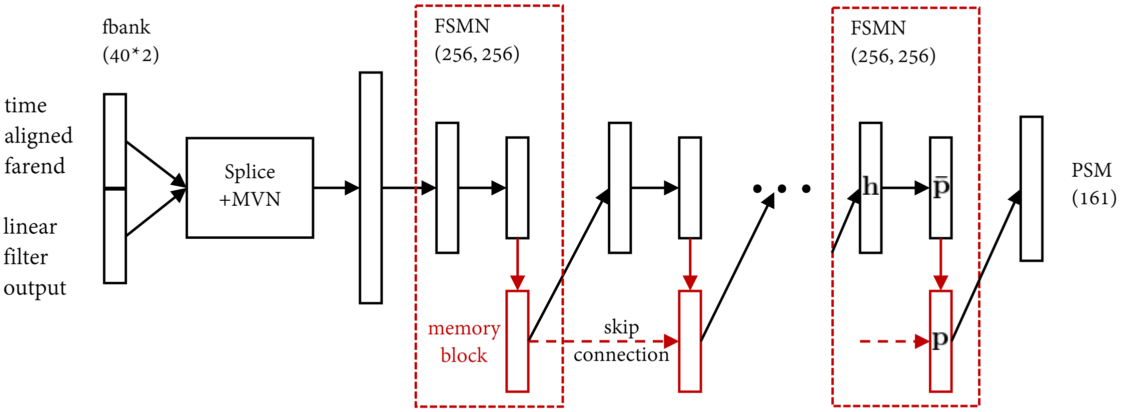

Fig. 2. The Deep-FSMN model for residual echo suppression.

2.3. Residual Echo Suppression

The Deep-FSMN model for residual echo suppression is illustrated in figure 2. Logarithm filter bank energies (fbank) of the time aligned far end and wRLS filter output signals are used as input to the neural network. The computation flow is given by:

$$\begin{aligned}
\mathbf{f}_{in} &= [\text{fbank}(\hat{S}_t), \text{fbank}(X_{t'})] \\
\mathbf{p}^1 &= \text{ReLU}(\mathbf{U^0}\mathbf{f}_{in} + \mathbf{v^0}) \\
\mathbf{p}^{j+1} &= \text{FSMN}(\mathbf{p}^j), \quad j \in [1, 2, ..., J-1] \\
\mathbf{f}_{out} &= \text{Sigmoid}(\mathbf{U}^{J+1}\mathbf{p}^J + \mathbf{v}^{J+1}) \qquad (15)
\end{aligned}$$

where $\mathbf{U}^j$ and $\mathbf{v}^j$ are respectively the weight matrix and bias vector in the j-th layer. Each FSMN block has one hidden layer, one projection layer and one memory block. The realization is given by:

$$\begin{aligned}
\mathbf{h}_t^j &= \text{ReLU}(\mathbf{U}_1^j\mathbf{p}_t^j + \mathbf{v}^j) \\
\bar{\mathbf{p}}_t &= \mathbf{U}_2^j\mathbf{h}_t^j \\
\mathbf{p}_t^{j+1} &= \mathbf{p}_t^j + \bar{\mathbf{p}}_t + \sum_{i=0}^{N} \mathbf{m}_i^j \odot \bar{\mathbf{p}}_{t-i} \qquad (16)
\end{aligned}$$

where $\mathbf{m}_i^j$ is a memory parameter weighting the history information $\bar{\mathbf{p}}_{t-i}$ and $\odot$ denotes element-wise multiplication. N is the look-back order. Skip connections are added between the memory blocks to alleviate the gradient vanishing problem in the training phase.

The training target is a modified version of the vanilla Phase Sensitive Mask (PSM) and is clipped to the range of [0,1]

$$\text{PSM} = \frac{|S_{t,f}|}{|\hat{S}_{t,f}|} \cdot \text{Re}(\frac{S_{t,f}}{\hat{S}_{t,f}}). \qquad (17)$$

Though complex masks as applied in the recent DNS-Challenge [18] have potentially better performance, no significant gains are observed in our preliminary experiments.

3. RELATION TO PRIOR WORK

Addressing AEC from the source separation perspective has been investigated in [19, 20], and ICA based solutions are discussed therein. Here, an Aux-ICA based solution is derived and results in a novel weighted RLS filter.

Exploiting deep neural networks for residual echo suppression is a trending practice in literature. Here we consider the capability of the causal Deep-FSMN architecture jointly with TDC and wRLS filter in a systematic view.

4. EXPERIMENTS

The AEC-Challenge dataset[1] covers the following scenarios: far end (FE) single talk (ST), with and without echo path change; near end (NE) single talk, no echo path change; double talk (DT), with and without echo path change. Both far and near end speech can be either clean or noisy. The evaluation is based on the P.808 Mean Opinion Score (MOS) [14] on a blind test set. The top 3 results are given in Table 1.

4.1. Algorithm Details

The wRLS adaptive filter is computed based on 20 ms frames with a hop size of 10 ms, and a 320-point discrete Fourier transform. A filter tap of $L = 5$ in Equation (4) is used, and the filter coefficients are updated as in Equation (11), with the correlation matrix $\mathbf{R}$ and correlation vector $\mathbf{r}$ estimated recursively using a smooth parameter of 0.8 and a source PDF shape parameter of $\beta = 0.2$ in Equation (14).

The TDC part is configured to cover a relative delay of up to 500 ms, which requires a 16384-point discrete Fourier transform. To reduce the computational complexity, the estimation is updated every 250 ms by Equation (3) and the calculation of $\Phi_{t,f}$ in different frequencies are spread evenly in this period.

[1] https://aec-challenge.azurewebsites.net/

Table 1. MOS across different test scenarios.

Team Id	ST NE MOS	ST FE Echo DMOS	DT Echo DMOS	DT Other DMOS
21	3.85	4.19	4.34	4.07
Ours	3.84	4.19	4.26	3.71
8	3.76	4.20	4.30	3.74
Baseline	3.79	3.84	3.84	3.28

For the residual echo suppression neural network, the inference process is computed as in Equation (15). The output $\mathbf{f}_{out}$ is point-wise multiplied with $\hat{S}_{t,f}$ for signal reconstruction. There are $J = 9$ FSMN blocks each with 256 hidden units, 256 projection units and a look-back order of $N = 20$. The input feature is a spliced by one frame in the past and one frame in the future, which leads to a vector dimension of 240, and then mean and variance normalized.

There are 1.4M trainable parameters in the model. The average time it takes to infer one frame is 0.61 ms (0.19 ms for TDC, wRLS and 0.42 ms for RES) on a Surface Laptop with Intel Core i5-8350U clocked at 1.9 GHz, based on an internal C++/SSE2 implementation.

4.2. Training Setup

For training the neural network, the first 500 clips in the official synthetic dataset are used as the validation set and the rest 9,500 utterances are used for training. Besides, the training data is augmented as follows:

1. Randomly remix the echo and near end speech in the official synthetic dataset (19,000 utterances).

2. Select far end single talk utterances in the real dataset and randomly remix with the near end speech (28,998 utterances).

3. Use sweep signals in the real dataset to estimate the echo paths and regenerate double talk data using utterances from the LibriSpeech corpus [21] with Signal-to-Echo Ratio (SER) uniformly distributed in [-6, 10] dB (25,540 utterances).

4. Regenerate 24,000 random room impulse responses in simulated rooms and selectively add audio effects [clipping, band-limiting, equalization, sigmoid-like transformation] to the echo signal (24,000 utterances).

The Deep-FSMN model is optimized using the Adam optimizer with a learning rate of 0.0003, under the mean squared error loss function. The model is first trained for 10 epochs on the 9,500 utterances, and then fine tuned on the augmented training set. The learning rate is decayed by 0.6 if the loss improvement is less than 0.001. The best model is selected based on the ITU-T recommendation P.862 Perceptual Evaluation of Speech Quality (PESQ) scores evaluated on the validation set.

4.3. Analysis

In Table 1, the baseline is a recurrent neural network that takes concatenated log power spectral features of the microphone signal and far end signal as input, and outputs a spectral suppression mask [13]. It performs reasonably well in the ST NE scenario, but lacks behind the top systems when echo exists. Informal listening indicates that our proposed algorithm sometimes over-suppresses the near end speech in double talk, which may explain the DT Other DMOS gap with the 1st system.

In Table 2, the proposed wRLS filter is compared with the linear filter in WebRTC-AEC3 [2] in terms of PESQ and Short-Time Objective Intelligibility (STOI) [22] on 500 clips of the validation set, and in terms of Echo Return Loss Enhancement (ERLE) on the ST FE in the test set. ERLE is defined as:

$$\text{ERLE} = 10 \log_{10} \frac{E[s^2(t)]}{E[\hat{s}^2(t)]} \tag{18}$$

Table 2. PESQ and STOI are evaluated on the synthetic validation set. ERLE is evaluated on the ST FE in the test set.

	PESQ	STOI	ERLE (dB)
Orig	1.24	0.79	-
WebRTC-AEC3	1.28	0.82	6.29
wRLS, $\beta = 0$	1.41	0.85	5.58
wRLS, $\beta = 0.2$	1.43	0.85	6.56
wRLS, $\beta = 0.4$	1.40	0.85	5.99
wRLS, $\beta = 1.0$	1.38	0.84	6.41
wRLS, $\beta = 0.2$ +Deep-FSMN	2.07	0.91	49.39

The performance of the wRLS filter varies with different source PDF shape parameters. A value of $\beta = 0.2$ is finally chosen, which outperforms AEC3 by 0.15 in PESQ, 0.03 in STOI and 0.27 dB in ERLE. The Deep-FSMN model greatly boost the overall performance, achieving a PESQ score of 2.07 and nearly complete echo reduction when echo exists.

5. CONCLUSION

This paper presents our submission to the AEC-Challenge. The algorithm achieves satisfactory subjective scores on real recordings by systematically combing time delay compensation, a novel wRLS linear filter and a Deep-FSMN model for residual echo suppression. The wRLS filter is derived from the semi-blind source separation reformulation of the acoustic echo cancellation problem and simplification of the Aux-ICA solution. One end-to-end neural network model that takes the raw near end mic signal and far end signal as input and outputs the near end speech is more appealing, which will be future direction of this work.

6. REFERENCES

[1] Jacob Benesty, Tomas Gänsler, Dennis R Morgan, M Mohan Sondhi, Steven L Gay, et al., "Advances in network and acoustic echo cancellation," 2001.

[2] webrtc, "https://webrtc.googlesource.com/src," .

[3] Charles Knapp and Glifford 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.

[4] Bjoern Voelcker and W Bastiaan Kleijn, "Robust and low complexity delay estimation," in *International Workshop on Acoustic Signal Enhancement*. VDE, 2012, pp. 1–4.

[5] Simon S Haykin, *Adaptive filter theory*, Pearson Education India, 2008.

[6] Chao Wu, Xiaofei Wang, Yanmeng Guo, Qiang Fu, and Yonghong Yan, "Robust uncertainty control of the simplified kalman filter for acoustic echo cancelation," *Circuits, Systems, and Signal Processing*, vol. 35, no. 12, pp. 4584–4595, 2016.

[7] John J Shynk et al., "Frequency-domain and multirate adaptive filtering," *IEEE Signal processing magazine*, vol. 9, no. 1, pp. 14–37, 1992.

[8] Hai Huang, Christian Hofmann, Walter Kellermann, Jingdong Chen, and Jacob Benesty, "A multiframe parametric wiener filter for acoustic echo suppression," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2016, pp. 1–5.

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

[10] Amin Fazel, Mostafa El-Khamy, and Jungwon Lee, "Deep multitask acoustic echo cancellation.," in *INTERSPEECH*, 2019, pp. 4250–4254.

[11] Hao Zhang, Ke Tan, and DeLiang Wang, "Deep learning for joint acoustic echo and noise cancellation with nonlinear distortions.," in *INTERSPEECH*, 2019, pp. 4255–4259.

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

[13] Ando Saabas Tanel Parnamaa Hannes Gamper Sebastian Braun Robert Aichner Sriram Srinivasan Kusha Sridhar, Ross Cutler, "Icassp 2021 acoustic echo cancellation challenge: Datasets and testing framework," *arXiv preprint arXiv:2009.04972*, 2020.

[14] Babak Naderi and Ross Cutler, "An open source implementation of itu-t recommendation p. 808 with validation," *arXiv preprint arXiv:2005.08138*, 2020.

[15] Shiliang Zhang, Ming Lei, Zhijie Yan, and Lirong Dai, "Deep-fsmn for large vocabulary continuous speech recognition," in *ICASSP*. IEEE, 2018, pp. 5869–5873.

[16] Nobutaka Ono and Shigeki Miyabe, "Auxiliary-function-based independent component analysis for super-gaussian sources," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2010, pp. 165–172.

[17] Ziteng Wang, Yueyue Na, Zhang Liu, Yun Li, Biao Tian, and Qiang Fu, "A semi-blind source separation approach for speech dereverberation," in *INTERSPEECH*, 2020.

[18] Chandan KA Reddy, Vishak Gopal, Ross Cutler, Ebrahim Beyrami, Roger Cheng, Harishchandra Dubey, Sergiy Matusevych, Robert Aichner, Ashkan Aazami, Sebastian Braun, et al., "The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," *Proc. Interspeech 2020*, pp. 2492–2496, 2020.

[19] Francesco Nesta, Ted S Wada, and Biing-Hwang Juang, "Batch-online semi-blind source separation applied to multi-channel acoustic echo cancellation," *IEEE transactions on audio, speech, and language processing*, vol. 19, no. 3, pp. 583–599, 2010.

[20] Ryu Takeda, Kazuhiro Nakadai, Toru Takahashi, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G Okuno, "Efficient blind dereverberation and echo cancellation based on independent component analysis for actual acoustic signals," *Neural computation*, vol. 24, no. 1, pp. 234–272, 2012.

[21] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[22] Tiago H Falk, Chenxi Zheng, and Wai-Yip Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1766–1774, 2010.