A NEW PERSPECTIVE OF AUXILIARY-FUNCTION-BASED INDEPENDENT COMPONENT ANALYSIS IN ACOUSTIC ECHO CANCELLATION

Yueyue Na, Ziteng Wang, Zhang Liu, Yun Li, Biao Tian, Qiang Fu

Machine Intelligence Technology, Alibaba Group
{yueyue.nyy, ziteng.wzt, yinan.lz, yl.yy, tianbiao.tb, fq153277}@alibaba-inc.com

ABSTRACT

For better speech communication ability, acoustic echo cancellation (AEC) is required to suppress echo from observed signals. Traditional AEC's performance is restricted when double-talk occurs. Blind source separation (BSS) based AEC has good performance in double-talking since BSS is inherently based on the full duplex signal mode, which both far-end and near-end signals coexist. In this paper, the auxiliary-function-based independent component analysis (Aux-ICA) algorithm is used to solve the AEC problem. After mathematical simplification, it can be seen that the Aux-ICA AEC is equivalent to weighted recursive least square (RLS) algorithm, with the ICA nonlinearity as the weighting function. The performance of the proposed approach is verified by both simulated and real-world experiments.

***Index Terms*—** acoustic echo cancellation, blind source separation, independent component analysis, double-talk

1. INTRODUCTION

Microphone array and loudspeakers are often installed on modern smart devices for better human-machine or human-human voice communication. In such cases, sound played by device itself will be collected back by its own microphone, i.e., the acoustic echo, which will contaminate the useful signal from real user. Thus, acoustic echo cancellation (AEC) technique is required to suppress echo from original signals.

Adaptive filtering [1] is a common way to perform AEC. Normalized least mean square (NLMS) approach [1-3] is a typical algorithm of this type, which minimizes the mean square error between echo estimation and microphone signal by gradient descent method. In order to prevent filter divergence, double-talk detector (DTD) [4], or adaptive step size policy [5] is used to freeze or slow down the filter adaptation when double-talk occurs. Recursive least square (RLS) [6, 7] is another class of AEC algorithm. Compared with NLMS, RLS has faster convergence speed but higher computational complexity (section 2.4 of [7]).

In addition to the linear echo, due to the nonlinear effect of the playback system, nonlinear echo also presents in real applications, which cannot be cancelled directly by linear adaptive filtering. Post filtering and acoustic echo suppression (AES) techniques, such as [8-10], are usually applied after adaptive filtering to reduce nonlinear echo. Deep learning based approaches, like [11, 12], are also developed for nonlinear echo suppression with the help of deep neural network's powerful nonlinear modeling ability.

To achieve better voice communication quality, not only acoustic echo, but also environmental noise must be handled. Blind source separation (BSS) is a technique which performs speech enhancement by separating sound components from observed noisy signals. Independent component analysis (ICA) [13, 14] and independent vector analysis (IVA) [15-19] are typical BSS techniques. AEC can also be considered as a semi-BSS problem which separates echo and near-end components from microphone signals, with references (far-end) as the supervised information. Researches about BSS based AEC can be found in [20-23], etc.

Traditional AEC adaptive filtering plus post filtering or AES techniques have some drawbacks. First, for gradient descent based approaches, there always exists the balance between convergence speed and stability [14]. Second, although DTD and adaptive step size policy can work well in single-talking and sparse double-talking scenes, their performance may degrade in continuous double-talking scene where near-end signals always exist [21, 23]. Third, most post filtering and AES approaches are based on Wiener filtering or masking techniques, like [10, 11], which suppress nonlinear echo at the price of near-end speech distortion.

Compared with traditional AEC approaches like [2, 3, 5-7], BSS based AEC algorithms have the advantage of good echo cancellation ability in continuous double-talking scene [23], since BSS signal model is a full duplex model, which both far-end and near-end signals coexist. In this paper, the auxiliary function based ICA (Aux-ICA) algorithm [14] is used to solve the AEC problem. In addition to the full duplex property, with the help of the auxiliary function technique, the explicit step size parameter is avoided, and the convergence speed is increased [14]. After mathematical simplification, it can be seen that the Aux-ICA AEC can be considered as a kind of weighted RLS algorithm, with the ICA nonlinearity as the weighting function. The following content is organized as follows. Section 2 introduces the signal model, section 3 depicts the algorithm derivation, experiments and comparisons are given in section 4, at last, section 5 gives the conclusion of this paper.

2. PROBLEM FORMULATION

2.1. Signal Model

The signal model used in this paper is depicted in Figure 1. The microphone signal x consists of two parts: the linear echo e and the near-end signal s, as shown in (1).

$$x = s + e = s + \sum_{m=1}^{R} a_m * r_m \tag{1}$$

Supposing there are R references in the system, e can be modeled as the sum of the convolution between individual reference $r_1, \cdots, r_R$ and the corresponding unknown echo path $a_1, \cdots, a_R$, where $*$ for convolution. The near-end signal s can be considered as the combination of N sources $s_1, \cdots, s_N$. Since nonlinear echo $\tilde{e}$ cannot be cancelled by linear AEC, it is also classified to the near-end component in Figure 1.

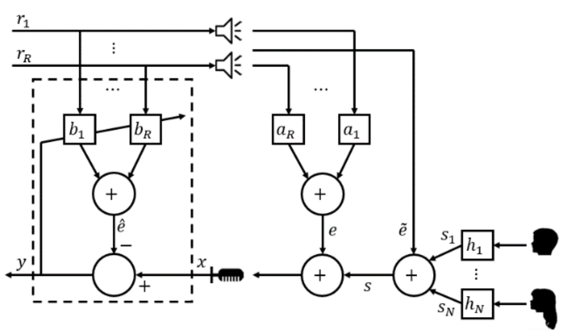

Fig. 1. Signal model.

AEC performs echo cancellation by calculating the echo estimation $\hat{e}$ from references and the estimated echo paths $b_1, \cdots, b_R$, then, subtracting $\hat{e}$ from x to get the output signal y, which is the estimation of the near-end signal s, as shown in (2), as well as the dotted box in Figure 1.

$$y = x - \hat{e} = x - \sum_{m=1}^{R} b_m * r_m \tag{2}$$

2.2. The NLMS and the RLS algorithm

Equation (3) and (4) depict the common NLMS iteration policy, where μ is the step size, δ is a small positive number to prevent zero denominator, $\boldsymbol{r}(\tau) = [r_1(\tau), \cdots, r_R(\tau - L + 1)]^T$ is reference vector with the length of LR, τ is frame index, L the number of frames [3, 7], ^{T} and ^{H} for transpose and conjugated transpose. In this paper, the estimated echo path is always denoted by $\boldsymbol{b}$, the vector length and conjugation is uniquely determined by the signal model context.

$$\boldsymbol{b}(\tau) = \boldsymbol{b}(\tau - 1) + \frac{\mu y^H(\tau)\boldsymbol{r}(\tau)}{\delta + \boldsymbol{r}^H(\tau)\boldsymbol{r}(\tau)} \tag{3}$$

$$y(\tau) = x(\tau) - \boldsymbol{b}^H(\tau)\boldsymbol{r}(\tau) \tag{4}$$

The RLS approach is based on the so called normal equation in (5), where $\boldsymbol{R}$ is the reference auto-correlation matrix, $\boldsymbol{p}$ is the reference-mic cross correlation vector, $0 \ll \alpha < 1$ is the forgetting factor [6, 7].

$$\boldsymbol{b}(\tau) = \boldsymbol{R}^{-1}(\tau)\boldsymbol{p}(\tau) \tag{5}$$

$$\boldsymbol{R}(\tau) = \alpha\boldsymbol{R}(\tau - 1) + (1 - \alpha)\boldsymbol{r}(\tau)\boldsymbol{r}^H(\tau) \tag{6}$$

$$\boldsymbol{p}(\tau) = \alpha\boldsymbol{p}(\tau - 1) + (1 - \alpha)\boldsymbol{r}(\tau)x^H(\tau) \tag{7}$$

Solving (5) directly is computationally intensive because of the matrix inversion operation. Several policies can be used to speed up the process, e.g., the Woodbury matrix identity [24], and the QR decomposition (chapter 3 of [7]), etc.

2.3. The BSS Model

The signal model in (1) and (2) can also be denoted as the BSS mixing and demixing model in (8) and (9), where $\boldsymbol{s} = [s, \boldsymbol{r}^T]^T$, $\boldsymbol{x} = [x, \boldsymbol{r}^T]^T$, and $\boldsymbol{y} = [y, \boldsymbol{r}^T]^T$ are $1 + LR$ dimension vectors [20-23].

$$\boldsymbol{x}(\tau) = \underbrace{\begin{bmatrix} 1 & a_1(\tau), \cdots, a_{LR}(\tau) \\ \boldsymbol{0}_{LR\times 1} & \boldsymbol{I}_{LR\times LR} \end{bmatrix}}_{\boldsymbol{A}(\tau)} \boldsymbol{s}(\tau) \tag{8}$$

$$\boldsymbol{y}(\tau) = \underbrace{\begin{bmatrix} 1 & b_1(\tau), \cdots, b_{LR}(\tau) \\ \boldsymbol{0}_{LR\times 1} & \boldsymbol{I}_{LR\times LR} \end{bmatrix}}_{\boldsymbol{B}(\tau)} \boldsymbol{x}(\tau) \tag{9}$$

The mixing matrix $\boldsymbol{A}$ and the demixing matrix $\boldsymbol{B}$ have their special structures, where $\boldsymbol{0}$ and $\boldsymbol{I}$ are zero and identity matrices with proper size [20, 21].

Unlike NLMS and RLS, s is explicitly modeled as an independent component in the BSS signal model. Thus, it can be considered as a full duplex model with both far-end and near-end signals coexist.

3. THE PROPOSED APPROACH

In this paper, the Aux-ICA algorithm [14, 17] is used to perform the separation. For the detailed theory and algorithm derivation of Aux-ICA (IVA), please refer to [14, 16-19].

3.1. Algorithm Derivation

According to Aux-ICA, the weighted correlation matrix $\boldsymbol{C}$ is updated first, as shown in (10) and (11), where $\varphi[\cdot]$ is a nonlinear transform. The nonlinearity in (11) [19] is used in this paper, where γ is the sparse parameter.

$$\boldsymbol{C}(\tau) = \alpha\boldsymbol{C}(\tau - 1) + \beta(\tau)\boldsymbol{x}(\tau)\boldsymbol{x}^H(\tau) \tag{10}$$

$$\begin{aligned} \beta(\tau) &= (1 - \alpha)\varphi[y(\tau)] \\ &= (1 - \alpha)(|y(\tau)|^2 + \delta)^{(\gamma-2)/2} \end{aligned} \tag{11}$$

Then, the demixing filter $\boldsymbol{w} = \boldsymbol{B}^H \boldsymbol{i}_1 = [1, \boldsymbol{b}^T]^T$ can be calculated according to (12) and (13), where $\boldsymbol{i}_1 = [1, 0, \cdots, 0]^T$ is a $1 + LR$ dimension vector. Unlike the source separation problem, there are no permutation and scaling ambiguities in the BSS based AEC. As a result, $\boldsymbol{w}$ is normalized according to (13) to make sure that the demixing matrix $\boldsymbol{B}$ maintains the structure in (9). After (13), $w_1 = 1$, and the rest elements of $\boldsymbol{w}$ stand for the estimated echo path $\boldsymbol{b}$.

$$\begin{aligned} \boldsymbol{w}(\tau) &= [\boldsymbol{B}(\tau-1)\boldsymbol{C}(\tau)]^{-1}\boldsymbol{i}_1 \\ &= \boldsymbol{C}^{-1}(\tau)\boldsymbol{B}^{-1}(\tau-1)\boldsymbol{i}_1 \end{aligned} \tag{12}$$

$$\boldsymbol{w} \leftarrow \boldsymbol{w}/w_1 \tag{13}$$

By noticing that $\boldsymbol{B}$ has the special structure in (9), $\boldsymbol{B}^{-1}$ must have the same structure. Thus, the inversion of $\boldsymbol{B}$ in (12) is avoided, as shown in (14).

$$\boldsymbol{w}(\tau) = \boldsymbol{C}^{-1}(\tau)\boldsymbol{i}_1 \tag{14}$$

The correlation matrix $\boldsymbol{C}$ in (10) and (12) can be partitioned according to (15) to (18).

$$\boldsymbol{C} = \begin{bmatrix} c_{11} & \boldsymbol{p}^H \\ \boldsymbol{p} & \boldsymbol{R} \end{bmatrix} \tag{15}$$

$$c_{11}(\tau) = \alpha c_{11}(\tau-1) + \beta(\tau)x(\tau)x^H(\tau) \tag{16}$$

$$\boldsymbol{p}(\tau) = \alpha\boldsymbol{p}(\tau-1) + \beta(\tau)\boldsymbol{r}(\tau)x^H(\tau) \tag{17}$$

$$\boldsymbol{R}(\tau) = \alpha\boldsymbol{R}(\tau-1) + \beta(\tau)\boldsymbol{r}(k)\boldsymbol{r}^H(\tau) \tag{18}$$

According to block matrix inversion, $\boldsymbol{C}^{-1}$ can be denoted in (19).

$$\boldsymbol{C}^{-1} = \begin{bmatrix} \dfrac{1}{c_{11} - \boldsymbol{p}^H\boldsymbol{R}^{-1}\boldsymbol{p}} & \dfrac{-\boldsymbol{p}^H}{c_{11}}\left(\boldsymbol{R} - \dfrac{\boldsymbol{p}\boldsymbol{p}^H}{c_{11}}\right)^{-1} \\ \dfrac{-\boldsymbol{R}^{-1}\boldsymbol{p}}{c_{11} - \boldsymbol{p}^H\boldsymbol{R}^{-1}\boldsymbol{p}} & \left(\boldsymbol{R} - \dfrac{\boldsymbol{p}\boldsymbol{p}^H}{c_{11}}\right)^{-1} \end{bmatrix} \tag{19}$$

Comparing (19) with (14) and (13), it is easy to derive the simplified solution in (20). The negative sign is discarded from (19) to (20) since the BSS demixing model in (9) can be interpreted as adding the negative echo estimation to microphone signals, rather than subtracting it as in traditional AEC model.

$$\boldsymbol{b}(\tau) = \boldsymbol{R}^{-1}(\tau)\boldsymbol{p}(\tau) \tag{20}$$

3.2. Discussion

The objective function of the RLS algorithm is minimizing the mean square error between echo estimation and microphone signals, and the corresponding optimization policy is the least square approach [6, 7]. Near-end signal is not explicitly modeled in the RLS signal model, and often be considered as noise in the system identification procedure. On the other hand, the objective function of Aux-ICA is minimizing the mutual information, which is measured by the KL divergence, and the auxiliary function technique is used for the optimization [14, 16-19]. Near-end signal is explicitly modeled as an independent component in the ICA model.

Although the basic theory of the two algorithms are different, from preceding subsection, it can be seen that the Aux-ICA AEC solution in (20) has the same form as the normal equation of the RLS algorithm in (5). However, from (6), (7) to (17), (18), the ICA nonlinearity β is used in the adaptation. Thus, the Aux-ICA AEC can be considered as a kind of weighted RLS algorithm, with the ICA nonlinearity as the weighting function. From the next section, it can be seen that the echo cancellation performance is improved by the nonlinear weighting.

As the conclusion to this section, the Aux-ICA AEC algorithm is summarized in Table 1. Fast implementation is not discussed here, readers can refer to [7] for further information.

Table 1. The Aux-ICA AEC algorithm.

Initialize: $\boldsymbol{b}(0) = \boldsymbol{0}$, $\boldsymbol{R}(0) = \boldsymbol{0}$, $\boldsymbol{p}(0) = \boldsymbol{0}$, $\alpha = 0.9999$, $\gamma = 0.2$, $\delta = 1 \times 10^{-10}$
Input: $x(\tau)$, $\boldsymbol{r}(\tau)$, output: $y(\tau)$
1. $y(\tau) = x(\tau) - \boldsymbol{b}^H(\tau-1)\boldsymbol{r}(\tau)$; 2. Calculate $\beta(\tau)$ according to (11); 3. Update $\boldsymbol{p}(\tau)$, $\boldsymbol{R}(\tau)$ according to (17) and (18); 4. Update $\boldsymbol{b}(\tau)$ according to (20);

4. EXPERIMENT

4.1. Simulated Experiment

A simulated environment is established: one speech and one music signal of 16k Hz sample rate are used as near-end and reference. Simulated data are generated according to (1), signal-to-echo ratio (SER) is adjusted to -20 dB. Echo paths of length 256 are randomly generated. Three algorithms are compared: NLMS, RLS, and Aux-ICA AEC (Aux). Filter size $L = 256$ for all algorithms, $\mu = 0.5$ in NLMS, which has balanced convergence speed and nearend distortion in this experiment. Misalignment in (21), and Perceptual Evaluation of Speech Quality (PESQ) [25] are used as performance indices. The simulated experimental environment is available at: https://github.com/nay0648/bssaec2020

$$Misalignment = 10\, log_{10}\left(\frac{\|\boldsymbol{b} - \boldsymbol{a}\|^2}{\|\boldsymbol{a}\|^2}\right) \tag{21}$$

Misalignment and corresponding near-end speech are shown in Figure 2. Several facts can be observed from Figure 2. First, the misalignment of NLMS and RLS is increased when double-talk occurs, however, the performance of Aux can keep relatively stable, which verifies the full duplex property of the Aux model. Second, Aux has the lowest misalignment and fastest convergence speed among the compared approaches. The average misalignment and PESQ are given in Table 2, which also supports the preceding observations.

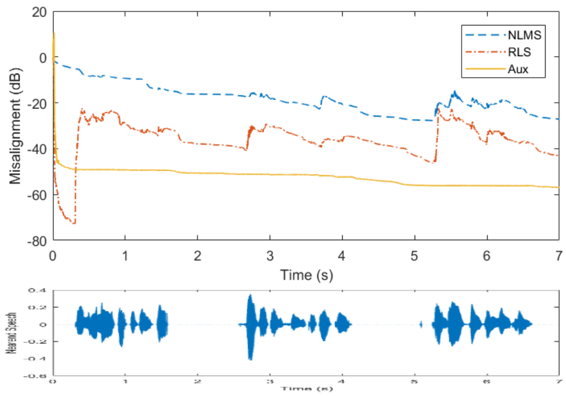

Fig. 2. Misalignment variation and near-end speech (colored).

Table 2. Average misalignment and PESQ.

	Average Misalignment (dB)	**PESQ**
NLMS	-18.15	1.17
RLS	-36.08	2.47
Aux	-52.31	4.12

4.2. Real-World Experiment

Real recordings are captured in a testing room configured in Figure 3. The smart TV with 4 mics and 4 loudspeakers is the sound capturing device, the captured signals are 16 kHz, 8 channels data, with 4 mic and 4 reference channels. The 4 mics are arranged into a linear array with 3.5 cm mic spacing, which is located at the bottom of the TV. Target and noise loudspeakers also exist in Figure 3. The loudspeakers marked as "Noise 1", "Noise 2", and "Subwoofer" are connected to the same audio hub for noise playback, the "Noise 1" and "Noise 2" loudspeakers are facing to the wall to increase the diffusion effect.

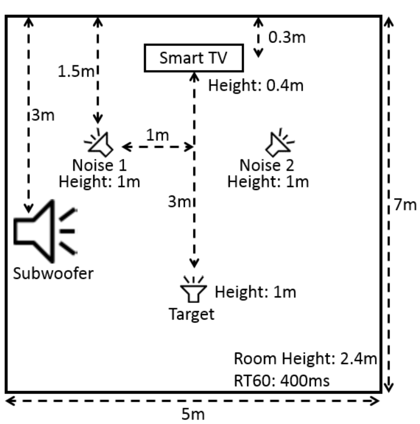

Fig. 3. Real-world experimental environment.

Target speech, echo, and noise are captured individually in the environment of Figure 3, then, added together as the algorithm input. The target source is 120 utterances of a certain keyword; the echo source is a music clip; the noise source is recorded in a TV store, which contains both diffuse noise and a news broadcast. The sound pressure level is tuned to 70, 90, and 75 dB(A) for target speech, echo, and noise, individually, according to a sound level meter placed close to the microphone array. Hence, the SER is -20 dB, and the signal-to-noise ratio (SNR) is -5 dB for the directly mixed signals.

The front-end used in this experiment utilizes the "AEC-(Aux-IVA [17])" architecture. The Aux-IVA part is fixed among all comparisons for source and nonlinear echo separation, while NLMS, RLS, and Aux are implemented in frequency domain as the "AEC" submodule. The FFT size and the frame shift is 640, 320 (not powers of 2 to compatible with the back-end system), $L = 10$ in all frequency bins. In order to perform objective comparison, the keyword spotting system introduced in [26] is concatenated to the front-end output. A keyword detected in any output channel is counted as one spot. Recall is used as the performance index, which is defined as: "number of spots / number of keywords".

Table 3 gives the keyword spotting results in the continuous double-talking scene with speech, echo, and noise coexist. In addition to direct mixing, target speech power is adjusted to generate different SER and SNR. First, it can be seen from Table 3 that the keyword can hardly be detected when the AEC submodule is bypassed in the front-end pipeline, which verifies the importance of AEC. It also can be seen that the performance of Aux is better than NLMS and RLS in this experiment, which is accordant with the conclusion of the last subsection.

Table 3. Recall in continuous double-talking scene.

	Recall (%)				
SER (dB)	**-25**	**-20**	**-19**	**-18**	**-17**
SNR (dB)	**-10**	**-5**	**-4**	**-3**	**-2**
AEC Bypass	1.7	0.8	1.7	1.7	0.8
NLMS	16.7	57.5	69.2	75.0	88.3
RLS	36.7	73.3	85.0	87.5	88.3
Aux	71.7	95.0	94.2	94.2	95.0

5. CONCLUSION

Unlike traditional AEC, e.g., NLMS and RLS, BSS AEC's signal model can be considered full duplex which both near-end and far-end signals coexist. Near-end speech is explicitly modeled as an independent component in the ICA model. Thus, the BSS AEC is expected to have better performance in continuous double-talking scene. In this paper, the Aux-ICA algorithm is used to solve the AEC problem. After mathematical simplification, it can be seen that the Aux-ICA AEC has strong connection with the RLS algorithm. Specifically, it can be considered as weighted RLS, with the ICA nonlinearity as the weighting function. The performance of the proposed approach is verified by both simulated and real-world experiments. In addition, a speech enhancement front-end is unified in the Aux-ICA/IVA framework for the real-world application of smart TV keyword spotting.

6. REFERENCES

[1] Haykin, Simon S. *Adaptive filter theory*. Pearson Education India, 2005.

[2] Morgan, Dennis R., and Steven G. Kratzer. "On a class of computationally efficient, rapidly converging, generalized NLMS algorithms." *IEEE Signal Processing Letters* 3.8 (1996): 245-247.

[3] Albu, Iuliana, Cristian Anghel, and Constantin Paleologu. "Adaptive filtering in acoustic echo cancellation systems—A practical overview." *2017 9th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*. IEEE, 2017.

[4] Yang, Jun. "Multilayer Adaptation Based Complex Echo Cancellation and Voice Enhancement." *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.

[5] Valin, Jean-Marc. "On adjusting the learning rate in frequency domain echo cancellation with double-talk." *IEEE Transactions on Audio, Speech, and Language Processing* 15.3 (2007): 1030-1034.

[6] Buchner, Herbert, Jacob Benesty, and Walter Kellermann. "Generalized multichannel frequency-domain adaptive filtering: efficient realization and application to hands-free speech communication." *Signal Processing* 85.3 (2005): 549-570.

[7] Apolinário, José Antonio, and R. Rautmann. *QRD-RLS adaptive filtering*. Ed. José Antonio Apolinário. New York: Springer, 2009.

[8] Franzen, Jan, and Tim Fingscheidt. "An efficient residual echo suppression for multi-channel acoustic echo cancellation based on the frequency-domain adaptive Kalman filter." *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.

[9] Hong, Jungpyo. "Stereophonic Acoustic Echo Suppression for Speech Interfaces for Intelligent TV Applications." *IEEE Transactions on Consumer Electronics* 64.2 (2018): 153-161.

[10] Huang, Hai, et al. "A multiframe parametric Wiener filter for acoustic echo suppression." *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2016.

[11] Lee, Chul Min, Jong Won Shin, and Nam Soo Kim. "DNN-based residual echo suppression." *Sixteenth Annual Conference of the International Speech Communication Association*. 2015.

[12] Carbajal, Guillaume, et al. "Multiple-input neural network-based residual echo suppression." *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.

[13] Hyvärinen, Aapo, and Erkki Oja. "Independent component analysis: algorithms and applications." *Neural networks* 13.4-5 (2000): 411-430.

[14] Ono, Nobutaka, and Shigeki Miyabe. "Auxiliary-function-based independent component analysis for super-Gaussian sources." *International Conference on Latent Variable Analysis and Signal Separation*. Springer, Berlin, Heidelberg, 2010.

[15] Kim, Taesu, et al. "Blind source separation exploiting higher-order frequency dependencies." *IEEE transactions on audio, speech, and language processing* 15.1 (2006): 70-79.

[16] Ono, Nobutaka. "Stable and fast update rules for independent vector analysis based on auxiliary function technique." *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2011.

[17] Taniguchi, Toru, et al. "An auxiliary-function approach to online independent vector analysis for real-time blind source separation." *2014 4th Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*. IEEE, 2014.

[18] Scheibler, Robin, and Nobutaka Ono. "Independent vector analysis with more microphones than sources." *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019.

[19] Ono, Nobutaka. "Auxiliary-function-based independent vector analysis with power of vector-norm type weighting functions." *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 2012.

[20] Nesta, Francesco, 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* 19.3 (2010): 583-599.

[21] Ikram, Muhammad Z. "Blind source separation and acoustic echo cancellation: A unified framework." *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012.

[22] Buchner, Herbert, and Walter Kellermann. "A fundamental relation between blind and supervised adaptive filtering illustrated for blind source separation and acoustic echo cancellation." *2008 Hands-Free Speech Communication and Microphone Arrays*. IEEE, 2008.

[23] Gunther, Jake. "Learning echo paths during continuous double-talk using semi-blind source separation." *IEEE transactions on audio, speech, and language processing* 20.2 (2011): 646-660.

[24] Woodbury, Max A. "Inverting modified matrices." *Memorandum report* 42.106 (1950): 336.

[25] Recommendation, ITU-T. "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs." *Rec. ITU-T P*. 862 (2001).

[26] Chen, Mengzhe, et al. "Compact Feedforward Sequential Memory Networks for Small-footprint Keyword Spotting." *Interspeech*. 2018.