A UNIFIED SPEECH ENHANCEMENT FRONT-END FOR ONLINE DEREVERBERATION, ACOUSTIC ECHO CANCELLATION, AND SOURCE SEPARATION

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

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

ABSTRACT

Dereverberation (DR), acoustic echo cancellation (AEC), and blind source separation (BSS) are the three most important submodules in speech enhancement front-end. In traditional systems, the three submodules work independently in a sequential manner, each submodule has its own signal model, objective function, and optimization policy. Although this architecture has high flexibility, the speech enhancement performance is restricted, since each submodule's optimum cannot guarantee the entire system's global optimum. In this paper, a unified signal model is derived to combine DR, AEC, and BSS together, and the online auxiliary-function based independent component/vector analysis (Aux-ICA/IVA) technique is used to solve the problem. The proposed approach has unified objective function and optimization policy, the performance improvement is verified by simulated experiments.

Index Terms— dereverberation, acoustic echo cancellation, blind source separation, independent component analysis, independent vector analysis

1. INTRODUCTION

In distant talking human-machine or human-human speech communication, competing interferences, ambient noise, acoustic echo, and room reverberation are main factors leading to signal-to-noise ratio (SNR) and target speech intelligibility degradation. In order to overcome these disadvantages, many speech enhancement algorithms are proposed.

Blind source separation (BSS) techniques, such as independent component/vector analysis (ICA/IVA) [1-8], is a common way to perform interference suppression by separating original sources from microphone signals. In addition to point sources, IVA can also be used to extract speech from ambient and diffuse noise background, i.e., the independent vector extraction (IVE) technique [9, 10].

Adaptive filtering theory [11], and acoustic echo cancellation (AEC) techniques are usually used to suppress acoustic echo from observed signals. Normalized least mean square (NLMS) [12, 13] is a classical AEC algorithm, which estimates the unknown echo path by gradient descent approach. Affine projection (AP) [14], and recursive least square (RLS) [15] are two other frequently used AEC techniques, with higher convergence speed but higher computational complexity [15].

Although IVA and AEC algorithms are originally designed for reverberant environments, large reverberation will increase model scale and computational complexity, and reduce separation performance, as well as convergence speed. Thus, dereverberation (DR) techniques are used to shorten late reverberation. Typical DR algorithm such as weighted prediction error (WPE) [16, 17], etc.

Interferences, noise, echo, and reverberation usually coexists in real applications. A single algorithm is not adequate to overcome such complicated situation, instead, a front-end system is required. Traditional front-end usually works in a sequential manner, with individual algorithms as its submodules, e.g., the AEC-DR-BSS, or Beamforming (BF)-AEC-DR structures [18, 19], etc. Such architecture is very flexible, different data flows and algorithms can be chosen for different applications. However, it also has some drawbacks. First, each algorithm is updated according to its own objective function, and converged to its own optimum. However, individual optimums not means the system's global optimum. Second, long FFT (e.g., 4096 for 16 kHz sample rate) is required by original IVA [5-8] to model long room impulse response (RIR). However, long FFT causes large output latency, which is not suitable for online applications.

Joint and unified algorithms are proposed to overcome the preceding drawbacks. E.g., a joint AEC and BF approach is proposed in [19] for better nonlinear echo suppression, a unified AEC and DR algorithm is proposed in [20], unified DR and BF algorithms are proposed in [21, 22]. Our previous works [23, 24] formulate AEC and DR under the auxiliary function based ICA (Aux-ICA) framework, this paper is a further step of [23, 24]. In this paper, the three submodules: DR, AEC, and BSS are formulated in a unified signal model, and the Aux-ICA/IVA technique is used to solve the problem. The separation performance of the proposed approach is improved compared to the simple concatenation of the three submodules. The rest of this paper is organized as follows: section 2 derives the signal model, section 3 depicts the algorithm, experiments and comparisons are given in section 4, at last, section 5 gives the conclusion of this paper.

2. SIGNAL MODEL

After short-time Fourier transform (STFT), time domain signals can approximately be converted to the signal model in (1), where $\boldsymbol{x} = [x_1, \cdots, x_M]^T$, $\boldsymbol{s} = [s_1, \cdots, s_N]^T$, and $\boldsymbol{r} = [r_1, \cdots, r_R]^T$ are microphone, source, and reference signals, M, N, and R are the number of microphones, sources, and references, ^{T} for transpose. Different from the long STFT model in [5-8], short STFT is used in (1) (e.g., 512 FFT size and 256 block shift for 16 kHz sample rate). Short STFT has lower output latency and balanced computational load, which is more suitable for online applications. As a result, one STFT block is not adequate to cover the full RIR, but only can model the direct and early reflection part [21, 22]. Thus, $\boldsymbol{A}$ and $\boldsymbol{B}$ are $M \times N$ and $M \times R$ matrices, which are the direct and early part of the source transfer function and echo path. $\boldsymbol{A}_l$ and $\boldsymbol{B}_l$ are the corresponding late reverberations, τ and l are STFT frame index and reverberation index. STFT frequency bin index is omitted to simplify the notation.

$$\boldsymbol{x}(\tau) = \boldsymbol{A}\boldsymbol{s}(\tau) + \boldsymbol{B}\boldsymbol{r}(\tau) + \sum_{l=1}^{\infty} [\boldsymbol{A}_l \boldsymbol{s}(\tau - l) + \boldsymbol{B}_l \boldsymbol{r}(\tau - l)] \quad (1)$$

Signal model in (1) can be converted to (2) (see Appendix), where $\boldsymbol{C}_l$ are $M \times M$ matrices. From (1) to (2) is one of the key points of this paper, which provides the possibility to focus AEC and BSS on the direct and early reverberation part of the signal, and left late reverberation to DR.

$$\boldsymbol{x}(\tau) = \boldsymbol{A}\boldsymbol{s}(\tau) + \boldsymbol{B}\boldsymbol{r}(\tau) + \sum_{l=1}^{\infty} \boldsymbol{C}_l \boldsymbol{x}(\tau - l) \quad (2)$$

Since reverberation decays with time, the infinite summation in (2) can be approximated by finite summation in (3) if L is large enough.

$$\boldsymbol{x}(\tau) \approx \boldsymbol{A}\boldsymbol{s}(\tau) + \boldsymbol{B}\boldsymbol{r}(\tau) + \sum_{l=1}^{L} \boldsymbol{C}_l \boldsymbol{x}(\tau - l) \quad (3)$$

Equation (3) can be converted to the matrix-vector multiplication form in (4), where $\overline{\boldsymbol{x}}(\tau) = [\boldsymbol{x}^T(\tau), \cdots, \boldsymbol{x}^T(\tau - L + 1)]^T$, $\boldsymbol{0}$ is zero matrix, $\boldsymbol{I}$ is identity matrix. To simplify the problem, the determined BSS model ($M = N$) is considered, thus, the matrix $\boldsymbol{P}$ in (4) is square.

$$\begin{bmatrix} \boldsymbol{x}(\tau) \\ \boldsymbol{r}(\tau) \\ \overline{\boldsymbol{x}}(\tau - 1) \end{bmatrix} = \underbrace{\begin{bmatrix} \boldsymbol{A}_{M\times N} & \boldsymbol{B}_{M\times R} & \boldsymbol{C}_{M\times ML} \\ \boldsymbol{0}_{(R+ML)\times N} & \boldsymbol{I}_{(R+ML)\times(R+ML)} \end{bmatrix}}_{\boldsymbol{P}_{(M+R+ML)\times(N+R+ML)}} \begin{bmatrix} \boldsymbol{s}(\tau) \\ \boldsymbol{r}(\tau) \\ \overline{\boldsymbol{x}}(\tau - 1) \end{bmatrix} \quad (4)$$

It is easy to verify that $\boldsymbol{P}$ is invertible if $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ are full rank, which is often satisfied in normal mixing conditions. This means that the demixing procedure can be modeled in (5), where $\boldsymbol{z} = [z_1, \cdots, z_N]^T$ is the estimated sources, $\boldsymbol{D}$, $\boldsymbol{E}$, and $\boldsymbol{F}$ are the estimated demixing matrix, echo path, and reverberation path, individually.

$$\begin{bmatrix} \boldsymbol{z}(\tau) \\ \boldsymbol{r}(\tau) \\ \overline{\boldsymbol{x}}(\tau - 1) \end{bmatrix} = \underbrace{\begin{bmatrix} \boldsymbol{D}_{N\times M} & \boldsymbol{E}_{N\times R} & \boldsymbol{F}_{N\times ML} \\ \boldsymbol{0}_{(R+ML)\times M} & \boldsymbol{I}_{(R+ML)\times(R+ML)} \end{bmatrix}}_{\boldsymbol{Q}_{(N+R+ML)\times(M+R+ML)}} \begin{bmatrix} \boldsymbol{x}(\tau) \\ \boldsymbol{r}(\tau) \\ \overline{\boldsymbol{x}}(\tau - 1) \end{bmatrix} \quad (5)$$

The demixing model in (5) meets the IVA signal model, which means that the IVA algorithm, such as [6], can be used to perform DR, AEC, and BSS in a unified framework. In addition, the special structure of $\boldsymbol{Q}$ means that only the first N rows need to be solved.

3. THE ALGORITHM

Although the unified framework has already achieved by (5), it is not the best way to solve it directly. For the source separation problem, full frequency band nonlinearity is required by IVA to overcome the permutation ambiguity. However, ICA which utilizes frequency bin-wise nonlinearity is adequate to solve DR and AEC [23, 24]. Thus, it's better to decompose the DR and AEC part from (5) for more detailed control of the nonlinearity calculation. In addition, since matrix operations usually have the complexity of squared order with respect to the matrix size, decomposing a large matrix into multiple smaller ones will increase convergence speed, stability, and reduce complexity for online algorithms.

3.1. Demixing Model Decomposition

The most directly decomposition of $\boldsymbol{Q}$ in (5) is (6), where the size of each $\boldsymbol{0}_i$ and $\boldsymbol{I}_j$ is uniquely determined by $\boldsymbol{D}$, $\overline{\boldsymbol{E}}$, $\overline{\boldsymbol{F}}$, and $\boldsymbol{Q}$. $\overline{\boldsymbol{E}}$ and $\overline{\boldsymbol{F}}$ have the same size with $\boldsymbol{E}$ and $\boldsymbol{F}$, there values are determined by $\boldsymbol{Q}_{draec} = \boldsymbol{Q}_{bss}^{-1}\boldsymbol{Q}$.

$$\boldsymbol{Q} = \underbrace{\begin{bmatrix} \boldsymbol{D} & \boldsymbol{0}_2 \\ \boldsymbol{0}_1 & \boldsymbol{I}_1 \end{bmatrix}}_{\boldsymbol{Q}_{bss}} \underbrace{\begin{bmatrix} \boldsymbol{I}_2 & \overline{\boldsymbol{E}} & \overline{\boldsymbol{F}} \\ \boldsymbol{0}_1 & & \boldsymbol{I}_1 \end{bmatrix}}_{\boldsymbol{Q}_{draec}} \quad (6)$$

Further decompositions are also available, e.g., (7) and (8), $\boldsymbol{Q}_{aec}$ and $\boldsymbol{Q}_{dr}$ are commutable because of their special structures.

$$\boldsymbol{Q} = \underbrace{\begin{bmatrix} \boldsymbol{D} & \boldsymbol{0}_2 \\ \boldsymbol{0}_1 & \boldsymbol{I}_1 \end{bmatrix}}_{\boldsymbol{Q}_{bss}} \underbrace{\begin{bmatrix} \boldsymbol{I}_2 & \overline{\boldsymbol{E}} & \boldsymbol{0}_3 \\ \boldsymbol{0}_1 & & \boldsymbol{I}_1 \end{bmatrix}}_{\boldsymbol{Q}_{aec}} \underbrace{\begin{bmatrix} \boldsymbol{I}_2 & \boldsymbol{0}_4 & \overline{\boldsymbol{F}} \\ \boldsymbol{0}_1 & & \boldsymbol{I}_1 \end{bmatrix}}_{\boldsymbol{Q}_{dr}} \quad (7)$$

$$\boldsymbol{Q} = \underbrace{\begin{bmatrix} \boldsymbol{D} & \boldsymbol{0}_2 \\ \boldsymbol{0}_1 & \boldsymbol{I}_1 \end{bmatrix}}_{\boldsymbol{Q}_{bss}} \underbrace{\begin{bmatrix} \boldsymbol{I}_2 & \boldsymbol{0}_4 & \overline{\boldsymbol{F}} \\ \boldsymbol{0}_1 & & \boldsymbol{I}_1 \end{bmatrix}}_{\boldsymbol{Q}_{dr}} \underbrace{\begin{bmatrix} \boldsymbol{I}_2 & \overline{\boldsymbol{E}} & \boldsymbol{0}_3 \\ \boldsymbol{0}_1 & & \boldsymbol{I}_1 \end{bmatrix}}_{\boldsymbol{Q}_{aec}} \quad (8)$$

Different demixing model decompositions not only tell us how the problem is solved, but also give the solving order.

E.g., (6) solves DR and AEC together, then BSS, (7) and (8) perform DR-AEC-BSS and AEC-DR-BSS. At the first glance, $\boldsymbol{Q}$ and the right hand side of (6) to (8) are equivalent. However, $\boldsymbol{Q}$ is unknown, and should be estimated from the observed signals. Thus, the estimation quality is heavily affected by the solving algorithm and solving order. E.g., (9) is also a possible decomposition, which performs BSS first. However, it is known from (1) that a single block is not enough to model the long convolution, the separation performance of (9) is restricted.

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{I}_2 & \boldsymbol{E} & \boldsymbol{F} \\ \boldsymbol{0}_1 & & \boldsymbol{I}_1 \end{bmatrix} \underbrace{\begin{bmatrix} \boldsymbol{D} & \boldsymbol{0}_2 \\ \boldsymbol{0}_1 & \boldsymbol{I}_1 \end{bmatrix}}_{\boldsymbol{Q}_{bss}} \tag{9}$$

In the next subsection, the algorithm corresponding to (6) is depicted, algorithms for (7) to (9) have similar structures as (6).

3.2. The Pseudo Code

The model in (6) can be considered as the two steps in (10) and (11):

$$\boldsymbol{y}(\tau) = \boldsymbol{x}(\tau) - \underbrace{[-\overline{\boldsymbol{E}} \quad -\overline{\boldsymbol{F}}]}_{\boldsymbol{G}} \begin{bmatrix} \boldsymbol{r}(\tau) \\ \overline{\boldsymbol{x}}(\tau-1) \end{bmatrix} \tag{10}$$

$$\boldsymbol{z}(\tau) = \boldsymbol{D}\boldsymbol{y}(\tau) \tag{11}$$

Aux-ICA [4] and Aux-IVA [5-8] are chosen to estimate $\boldsymbol{G}$ and $\boldsymbol{D}$ iteratively. The work in [23] tells that solving (10) by Aux-ICA can be simplified as a weighted RLS algorithm, with the ICA nonlinearity as the weighting function. The minus sign in (10) is inherited from the classical AEC theory, which means subtracting echo from microphone signals.

The DRAEC-BSS algorithm corresponding to (10) and (11) is given in Table 1, where α is the forgetting factor, δ is used to prevent zero denominator, and γ is the nonlinearity's sparsity parameter, there values are tuned then fixed in all experiments. The nonlinearity (13) [8] is used in this paper, however, other nonlinearities also can be chosen. Equation (16) contains matrix inversion, which is not suitable for online applications, instead, the IQRD-RLS algorithm [15] can be used to reduce complexity. The derivation of the IQRD-RLS implementation of the DRAEC-BSS algorithm is out the scope of this paper, readers can refer to section 3.6 in [15] for more information.

3.3. Differences with Prior Works

Although (6) to (8) still use sequential process denoted by matrix multiplications, they are derived from the unified signal model in (2), which can be solved by the same optimization technique, i.e., Aux-ICA/IVA. In addition, AEC and BSS with single filter length are enabled from the derived signal model, which is suitable for online applications. The work in [20] unifies DR and AEC in ICA framework, which does the similar job as the DRAEC part of the proposed algorithm. However, the signal models are different. Speech, reference, and their histories are modeled in [20], while current speech, reference, and historical microphone signals are modeled in DRAEC. The derivation from (1) to (2) shows that the information of source and reference history is already contained in the historical microphone signals, so, the redundant information is removed in (2). DRAEC is also similar with WPE [16, 17] and the work in [24]. The difference is that current reference block is added to the signal model to perform AEC and DR together. In addition, according to (2), there is no delayed blocks between current and historical information in DRAEC, while, the delayed blocks is usually required in WPE to protect speech signals.

Table 1. The DRAEC-BSS algorithm.

Initialize:
$\boldsymbol{G}(0) = \boldsymbol{0}$, $\boldsymbol{\Phi}_{xr}(0) = \boldsymbol{0}$, $\boldsymbol{\Phi}_{rr}(0) = \boldsymbol{0}$, $\boldsymbol{D}(0) = \boldsymbol{I}$, $\alpha = 0.999$, $\delta = 0.01$, $\gamma = 0.2$

Input: $\boldsymbol{x}(\tau)$, $[\boldsymbol{r}^T(\tau), \overline{\boldsymbol{x}}^T(\tau-1)]^T$, output: $\boldsymbol{z}(\tau)$

1. Calculate dereverbed nearend signals as (12):

$$\boldsymbol{y}(\tau) = \boldsymbol{x}(\tau) - \boldsymbol{G}(\tau-1) \begin{bmatrix} \boldsymbol{r}(\tau) \\ \overline{\boldsymbol{x}}(\tau-1) \end{bmatrix} \tag{12}$$

2. Calculate ICA nonlinearity as (13):

$$\beta(\tau) = (1-\alpha)(\|\boldsymbol{y}(\tau)\|^2 + \delta)^{(\gamma-2)/2} \tag{13}$$

3. Update weighted mic-reference cross correlation as (14), ^{H} for conjugate transpose.

$$\boldsymbol{\Phi}_{xr}(\tau) = \alpha\boldsymbol{\Phi}_{xr}(\tau-1) + \beta(\tau)\boldsymbol{x}(\tau) \begin{bmatrix} \boldsymbol{r}(\tau) \\ \overline{\boldsymbol{x}}(\tau-1) \end{bmatrix}^H \tag{14}$$

4. Update weighted reference auto-correlation as (15):

$$\begin{aligned} &\boldsymbol{\Phi}_{rr}(\tau) \\ &= \alpha\boldsymbol{\Phi}_{rr}(\tau-1) \\ &+ \beta(\tau) \begin{bmatrix} \boldsymbol{r}(\tau) \\ \overline{\boldsymbol{x}}(\tau-1) \end{bmatrix} \begin{bmatrix} \boldsymbol{r}(\tau) \\ \overline{\boldsymbol{x}}(\tau-1) \end{bmatrix}^H \end{aligned} \tag{15}$$

5. Update DRAEC filters as (16):

$$\boldsymbol{G}(\tau) = \boldsymbol{\Phi}_{xr}(\tau)\boldsymbol{\Phi}_{rr}^{-1}(\tau) \tag{16}$$

6. The BSS part:

$$\boldsymbol{z}(\tau) = \boldsymbol{D}(\tau-1)\boldsymbol{y}(\tau) \tag{17}$$

Then, update $\boldsymbol{D}(\tau-1)$ to $\boldsymbol{D}(\tau)$ according to [5-8].

4. EXPERIMENT

4.1. Experimental Environment

A simulated room with the size of 5 by 7 by 2.4 m is established, the reverberation time (RT60) is 400 ms. $M = N = 2$, and $R = 1$. Sources and microphone array positions, as well as array orientations are randomly generated, the mic spacing is 10 cm. A simulated loudspeaker is placed at 10 cm below

the array. RIRs are simulated according to the image method [25, 26]. Target, interference, and reference are 24 seconds, 16 kHz sample rate data. Signals are convolved with the simulated RIRs, then, mixed according to SNR equals to 0 dB, and signal-to-echo ratio (SER) equals to 0 dB.

Six approaches are compared: WPE [16]-NLMS [13]-BSS, NLMS-WPE-BSS, DR-AEC-BSS in (7), AEC-DR-BSS in (8), DRAEC-BSS in (6), and BSS in (5). FFT size and STFT frame shift is 512/256, $L = 5$ for the DR and WPE part, the delay between current and historical blocks is set to 2 for WPE. Signal-to-distortion ratio (SDR) [27] improvement is used as the performance index. Except the open source WPE, all algorithm iterations are online, which means that the performance index is also affected by convergence speed and stability. The experimental environment is available at: https://github.com/nay0648/unified2021

4.2. Result and Discussion

The boxplot of 100 experiments is shown in Fig. 1. Several facts can be observed from the simulated experiments. First, WPE-NLMS-BSS and NLMS-WPE-BSS have restricted SDR improvements, which reveals that simple concatenation of submodules with different objective functions and optimization policies may not achieve the best performance. Second, although the BSS approach in (5) has unified objective function and optimization policy, the performance is still low, which verifies the analysis at the beginning of section 3. Third, DRAEC-BSS has the best performance among the compared approaches.

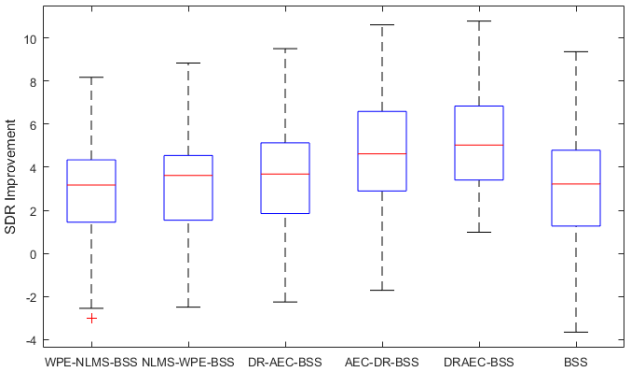

Fig. 1. SDR improvement.

The reason why the DRAEC-BSS approach has good performance can be deduced as follows. First, the unified signal model, objective function, and optimization policy may gain extra profit than the separated approaches. Second, although the same Aux-ICA/IVA technique is applied to the last four approaches, their nonlinearities are different. The DRAEC nonlinearity is calculate from the dereverbed near-end signals, which is more close to the required signals. On the other hand, in DR-AEC-BSS, the nonlinearity of the DR part is calculated from the dereverbed signals, which contain both nearend and echo. The existing echo may affect the quality of the nonlinearity. The DR part of the AEC-DR-BSS approach also calculates nonlinearity from the dereverbed near-end signals, however, without the help of DR, relatively long echo path may not be modeled well by single filter length AEC. The weakness of the BSS (5) nonlinearity is already analyzed at the beginning of section 3. Third, thanks to the signal model in (2), the BSS part (17) with short STFT blocks can be used. Compared with the long STFT BSS, online BSS with short block size has faster convergence speed, since less information need to be processed in one iteration.

5. CONCLUSION

In this paper, DR, AEC, and BSS are combined in a single front-end system. The proposed system has unified signal model with short STFT blocks, which is suitable for online applications. Weighted RLS, which is derived from Aux-ICA, then Aux-IVA are applied iteratively to perform speech enhancement. The unified architecture makes the proposed approach more likely to find better solution than traditional systems with separated submodules. The performance of the proposed approach is verified by simulated experiments. The algorithm in Table 1 is used to depict the theory, for real applications, the IQRD-RLS approach [15] should be used instead of the matrix inversion in (16).

APPENDIX

The proof from (1) to (2). First, expanding (2) to (a1):

$$\boldsymbol{x}(\tau) = \boldsymbol{A}\boldsymbol{s}(\tau) + \boldsymbol{B}\boldsymbol{r}(\tau) + \boldsymbol{C}_1\boldsymbol{x}(\tau-1) + \boldsymbol{C}_2\boldsymbol{x}(\tau-2) + \cdots \tag{a1}$$

Equation (a1) also holds at the $\tau - 1$ time instant in (a2):

$$\boldsymbol{x}(\tau-1) = \boldsymbol{A}\boldsymbol{s}(\tau-1) + \boldsymbol{B}\boldsymbol{r}(\tau-1) + \boldsymbol{C}_1\boldsymbol{x}(\tau-2) + \boldsymbol{C}_2\boldsymbol{x}(\tau-3) + \cdots \tag{a2}$$

Substituting (a2) to (a1) yields:

$$\boldsymbol{x}(\tau) = \boldsymbol{A}\boldsymbol{s}(\tau) + \boldsymbol{B}\boldsymbol{r}(\tau) + \boldsymbol{C}_1[\boldsymbol{A}\boldsymbol{s}(\tau-1) + \boldsymbol{B}\boldsymbol{r}(\tau-1) + \boldsymbol{C}_1\boldsymbol{x}(\tau-2) + \boldsymbol{C}_2\boldsymbol{x}(\tau-3) + \cdots] + \boldsymbol{C}_2\boldsymbol{x}(\tau-2) + \cdots \tag{a3}$$

It's easy to find that (a3) can be reformulated to (a4), where $\boldsymbol{A}_1 = \boldsymbol{C}_1\boldsymbol{A}$, $\boldsymbol{B}_1 = \boldsymbol{C}_1\boldsymbol{B}$, $\bar{\boldsymbol{C}}_2 = \boldsymbol{C}_1\boldsymbol{C}_1 + \boldsymbol{C}_2$, $\bar{\boldsymbol{C}}_3 = \boldsymbol{C}_1\boldsymbol{C}_2 + \boldsymbol{C}_3$, …

$$\boldsymbol{x}(\tau) = \boldsymbol{A}\boldsymbol{s}(\tau) + \boldsymbol{B}\boldsymbol{r}(\tau) + \boldsymbol{A}_1\boldsymbol{s}(\tau-1) + \boldsymbol{B}_1\boldsymbol{r}(\tau-1) + \bar{\boldsymbol{C}}_2\boldsymbol{x}(\tau-2) + \bar{\boldsymbol{C}}_3\boldsymbol{x}(\tau-3) + \cdots \tag{a4}$$

Repeating preceding steps at $\tau - 2$, $\tau - 3$, …, the final result is equation (1), which means that (1) and (2) are equivalent.

6. REFERENCES

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

[2] 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.

[3] Lee, Intae, Taesu Kim, and Te-Won Lee. "Fast fixed-point independent vector analysis algorithms for convolutive blind source separation." *Signal Processing* 87.8 (2007): 1859-1871.

[4] 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.

[5] 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.

[6] 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.

[7] Ono, Nobutaka. "Fast stereo independent vector analysis and its implementation on mobile phone." *IWAENC* 2012; International Workshop on Acoustic Signal Enhancement. VDE, 2012.

[8] 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.

[9] Ikeshita, Rintaro, Tomohiro Nakatani, and Shoko Araki. "Overdetermined independent vector analysis." *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020.

[10] Scheibler, Robin, and Nobutaka Ono. "Fast independent vector extraction by iterative SINR maximization." *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020.

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

[12] Slock, Dirk TM. "On the convergence behavior of the LMS and the normalized LMS algorithms." *IEEE Transactions on Signal Processing* 41.9 (1993): 2811-2825.

[13] 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.

[14] Gay, Steven L. "The fast affine projection algorithm." *Acoustic signal processing for telecommunication*. Springer, Boston, MA, 2000. 23-45.

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

[16] Nakatani, Tomohiro, et al. "Speech dereverberation based on variance-normalized delayed linear prediction." *IEEE Transactions on Audio, Speech, and Language Processing* 18.7 (2010): 1717-1731.

[17] Yoshioka, Takuya, and Tomohiro Nakatani. "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening." *IEEE Transactions on Audio, Speech, and Language Processing* 20.10 (2012): 2707-2720.

[18] Yang, Jun. "Multilayer adaptation based complex echo cancellation and voice enhancement." *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.

[19] Cohen, Alejandro, et al. "Joint beamforming and echo cancellation combining QRD based multichannel aec and MVDR for reducing noise and non-linear echo." *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018.

[20] Takeda, Ryu, et al. "ICA-based efficient blind dereverberation and echo cancellation method for barge-in-able robot audition." *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009.

[21] Nakatani, Tomohiro, and Keisuke Kinoshita. "Simultaneous Denoising and Dereverberation for Low-Latency Applications Using Frame-by-Frame Online Unified Convolutional Beamformer." *INTERSPEECH*. 2019.

[22] Boeddeker, Christoph, et al. "Jointly optimal dereverberation and beamforming." *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020.

[23] Yueyue Na, et al. "A New Perspective of Auxiliary-Function-Based Independent Component Analysis in Acoustic Echo Cancellation." https://github.com/nay0648/bssaec2020

[24] Ziteng Wang, et al. "A Semi-blind Source Separation Approach for Speech Dereverberation." *INTERSPEECH*. 2020.

[25] Allen, Jont B., and David A. Berkley. "Image method for efficiently simulating small‐room acoustics." *The Journal of the Acoustical Society of America* 65.4 (1979): 943-950.

[26] Habets, Emanuel AP. "Room impulse response generator." *Technische Universiteit Eindhoven, Tech. Rep* 2.2.4 (2006): 1.

[27] Vincent, Emmanuel, Rémi Gribonval, and Cédric Févotte. "Performance measurement in blind audio source separation." *IEEE transactions on audio, speech, and language processing* 14.4 (2006): 1462-1469.