A new acoustic echo cancellation framework combined with blind source separation

Renjie Wu[a, *], Lie Chen[b], Jucai Lin[c], Jun Yin[d]

ZheJiang Dahua Technology CO., LTD, Hangzhou 310000, China

[a]garrywrj@126.com, [b]195517048@qq.com, [c]lin_jucai@dahuatech.com, [d]yin_jun@dahuatech.com

*Corresponding author

Keywords: acoustic echo cancellation, double-talk, blind source separation.

Abstract: Acoustic echo cancellation has conventionally employed all variants known from deterministic adaptive filter design. The presence of double talk makes adaptive algorithm divergent. Meanwhile, a double talk detector with high accuracy and low complexity cannot be implemented easily. In this paper, we explore interesting connections between blind source separation and acoustic echo cancellation, and develop a framework which blind source separation separates mixed signal in double-talk scenario to avoid algorithm diverge instead of double talk detector. The forward BSS is employed as a preprocessor to separate near-end speech while AEC cancels the residual echo. The simulation results are evaluated with ERLE and its performance shows that the proposed framework is effective in double talk scenario.

1. Introduction

With the development of audio conference and hands-free technology, the communication between human to human and human to machine becomes more and more convenient. However, echo caused by acoustic coupling between the microphone and the loudspeaker ruins the communication quality when using hands-free devices. Therefore, an acoustic echo canceller is strongly required to estimate the echo signal to cancel the echo. A common AEC consists of an adaptive filter to estimate the acoustic impulse response of the near-end room, and uses it to produce a replica of the echo. The replica of echo is subtracted from the near-end microphone signal to send an echo-free signal to the far-end.

Nonetheless, the presence of near-end speech and noise make adaptive algorithm become unstable and divergent. Some double-talk detectors (DTD) are implemented to freeze filter adaptation in the double-talk scenario to avoid the divergence. However, the DTD has its own inherent disadvantage. It cannot give a precise estimation of start and end of the near-end speech. So the acoustic impulse response may change when the adaptation is frozen [1, 2].

In this paper, we propose a new framework to cancel the acoustic echo without the DTD. In this framework, Blind source separation (BSS) technology is combined to separate desired speech from mixed near-end microphone signal to avoid algorithm diverge in double-talk scenario and an adaptive volterra filter is implemented to cancel the residual echo. In this work, we will show how two channel BSS combined AEC can be used in both single and double talk scenario. The proposed framework achieves a good performance for AEC in comparison with the conventional AEC system. The organization of the paper is as follows: in Section 2, we will give an overview of conventional approach for AEC. The proposed framework will be described in detail in Section 3. Section 4 compares results in cancelling echo with ERLE criterion, followed by the concluding remarks presented in Section 5.

Published by CSP © 2019 the Authors

2. Conventional AEC systems

The structure of conventional AEC is shown in Fig. 1. The far-end signal *x*(*n*) is transmitted to the near-end room through the line or network, the echo signal *y*(*n*) is the result of convolution of far-end signal *x*(*n*) and the room impulse response (RIR), the microphone signal *d*(*n*) can be expressed as

$$d(n) = y(n) + v(n) = \sum_{l=0}^{L-1} h_l x(n-l) + v(n) = h^T x(n) + v(n) \tag{1}$$

Where $h = \begin{bmatrix} h_0 & \cdots & h_{L-1} \end{bmatrix}^T$ the coefficient vector for the room impulse response is, $x(n) = \begin{bmatrix} x(n) & \cdots & x(n-L+1) \end{bmatrix}^T$ is the vector for the far-end signal x (n), v (n) is the near-end speech.

The classical NLMS algorithm for AEC uses its adaptive filter to approximate room impulse response to get $\hat{h}$ and replica echo $\hat{y}(n)$ as

$$\hat{y}(n) = \hat{h}^T x(n) \tag{2}$$

Which is subtracted from the microphone signal to obtain error signal *e* (*n*) as

$$e(n) = y(n) - \hat{y}(n) \tag{3}$$

And the rule of coefficient iteration can be summarized as

$$\hat{h}(n+1) = \hat{h}(n) + \Delta\hat{h}(n) \tag{4}$$

$$\Delta\hat{h}(n) = \frac{\alpha}{x^T(n)x(n) + \beta} e(n)x(n) \tag{5}$$

Where *α* is step size, *β* is very small positive constant to avoid division by the zero. The result of the above adaptive process are as follows: $\hat{h}(n) \to h(n)$, $e(n) = y(n) - \hat{y}(n) \to 0$. And the DTD controls adaptation freeze or not through signal *x* (*n*), *d* (*n*) and *e* (*n*) to deal with the scene of double talk. However, a DTD with high accuracy and low complexity cannot be implemented easily.

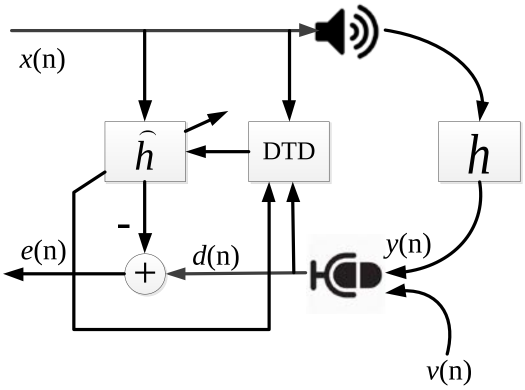

Figure 1. Conventional framework of AEC

3. Development of the proposed framework

In order to avoid disturbance of the estimation of room impulse response, a DTD-free framework has been proposed and its structure is shown in Fig. 2. Firstly, a two-channel microphone collects the mixed signal $m_1(n)$ and $m_2(n)$ in near-end room including far-end signal *x* (*n*) and near-end speech *v* (*n*). Then blind source separation is adopted to separate mixed signal to get the signal $d_1(n)$ which is close to *v*(*n*) and still a part of echo needed to be cancelled. Finally, adaptive filter cancel the residual echo to send the echo-free signal to the far-end.

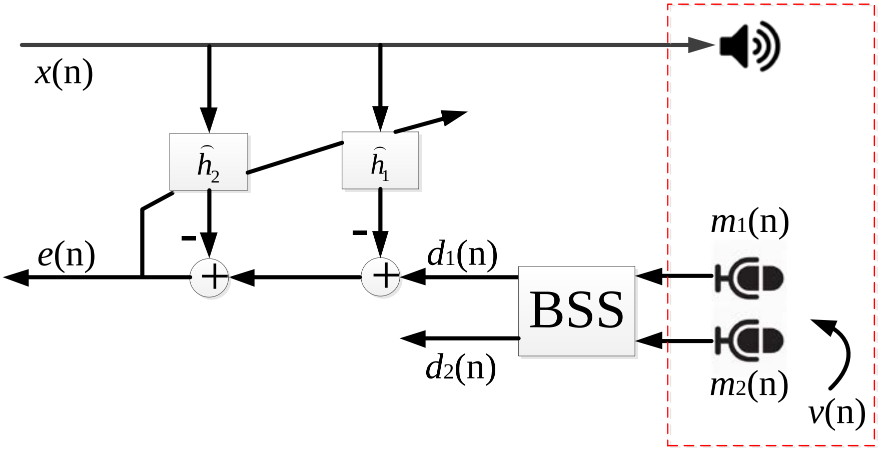

Figure 2. Descriptive scheme of the proposed framework

3.1 Blind source separation

BSS is a method for recovering a set of signals from the observation of their mixtures without any prior knowledge about the mixing process [3, 4], which is quite reliable to AEC preprocessing. A two-channel microphone is used in far-end room to capture the spatial information of the mixed signal, which is a great difference from conventional AEC framework. As shown in the Fig. 2, mixed signal at the microphone can be expressed as

$$\begin{bmatrix} m_1(n) \\ m_2(n) \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix} \begin{bmatrix} x(n) \\ v(n) \end{bmatrix} \tag{6}$$

Where x (n), v (n) is far-end signal and near-end speech, $H = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix}$ is indicated as the mixing matrix. The separated signal can be estimated by linearly demixing as

$$\begin{bmatrix} d_1(n) \\ d_2(n) \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix} \begin{bmatrix} m_1(n) \\ m_2(n) \end{bmatrix} \tag{7}$$

Where $W = \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix}$ is a demixing matrix which can be estimated through a gradient adaption as

$$W = W + \Delta W \tag{8}$$

Where ΔW is the gradient, which takes different forms according to the cost function that is to be minimized. Adopting the natural gradient adaption based on the minimization of the Kullback-Leibler divergence, ΔW is determined as

$$\Delta W = (I - E[\phi(d)d^T])W \tag{9}$$

Where $E[\cdot]$ is the expectation operator. The non-linearity $\phi(d)$ is determined from the expected probability density function of the output sources. A non-linearity with super-gaussian source is

$$\phi(d) = 2\tanh d \tag{10}$$

In ideal conditions, it can be seen that $d_1(n)$ will be approximate to near-end speech after several iterations but residual echo still exists, which should be cancelled by backward AEC algorithm.

3.2 Acoustic echo cancellation

In this section, an AEC based on forward BSS is explained. As the BSS method sending the signal $d_1(n)$ to AEC, a single channel adaptive volterra filter is used to cancel the far-end echo. For real world echo cancellation, a linear approximation may not achieve good performance [5] due to the nonlinear interference. A second order volterra filter [6] is used and mathematically described as

$$\hat{y}(n) = \hat{h}_0 + \hat{h}_1^T(n)x_1(n) + \hat{h}_2^T(n)x_2(n) \tag{11}$$

$$e(n) = d_1(n) - \hat{y}(n) \tag{12}$$

Where

$$\hat{h}_1(n) = [\hat{h}(0) \quad \cdots \quad \hat{h}(L-1)]^T \tag{13}$$

$$x_1(n) = [x(n) \quad \cdots \quad x(n-L+1)]^T \tag{14}$$

$$\hat{h}_2(n) = [\hat{h}(0,0) \quad \hat{h}(0,1) \quad \cdots \quad \hat{h}(0,L-1) \quad \hat{h}(1,1) \quad \cdots \quad \hat{h}(L-1,L-1)]^T \tag{15}$$

$$x_2(n) = [x^2(n) \quad x(n)x(n-1) \quad \cdots \quad x(n)x(n-L+1) \quad x^2(n-1) \quad \cdots \quad x^2(n-L+1)]^T \tag{16}$$

Where $\hat{h}_1(n)$, $\hat{h}_2(n)$ are first and second order filter tap weights, L is the length of the tap weights. So the length of coefficients in first kernel is L and second kernel has L (L+1)/2 number of coefficients. The volterra filter output $\hat{y}(n)$ is the sum of first and second order kernels and the weight of filter updating is done as given in (17)-(19).

$$\hat{h}_0(n+1) = \hat{h}_0(n) + \mu_1 e(n) \tag{17}$$

$$\hat{h}_1(n+1)=\hat{h}_1(n)+\mu_1 e(n)x_1(n) \tag{18}$$

$$\hat{h}_2(n+1)=\hat{h}_2(n)+\mu_2 e(n)x_2(n) \tag{19}$$

Where μ_0, μ_1 and μ_2 are different step size for volterra kernels.

4. Simulation results

The simulation is carried out to demonstrate the effectiveness of the proposed framework. The performance is measured by Echo return loss enhancement (ERLE). ERLE is the ratio of input mixed signal power to the power of a residual error signal immediately after blind source separation and echo cancellation and measured in dB.

$$ERLE=10\log_{10}\frac{E[m_1^2(n)]}{E[e^2(n)]} \tag{20}$$

In our experiment, a male speaker is at the far end of the conversation whereas a female speaker is at the near end. The two people speak to each other to ensure a double-talk scenario. The far-end signal and near-end microphone signal are given in Fig. 3, Fig. 4 and Fig. 5.

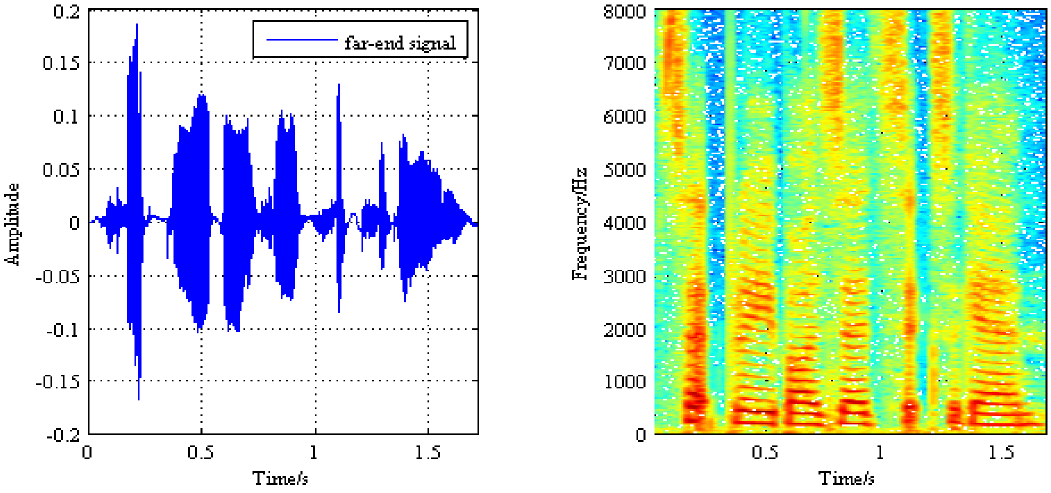

Figure 3. Time evolution and spectrograms of far-end signal x (n)

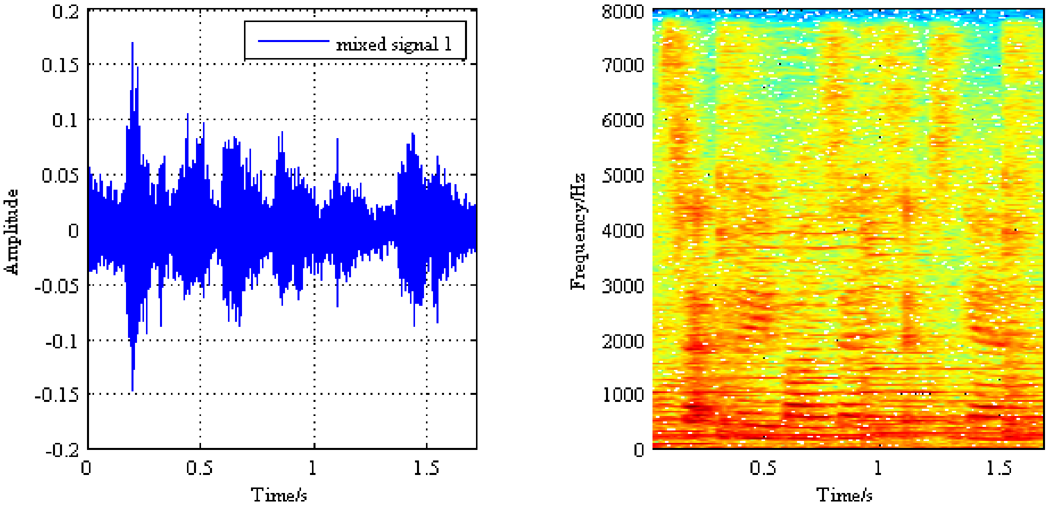

Figure 4. Time evolution and spectrograms of mixed signal $m_1(n)$

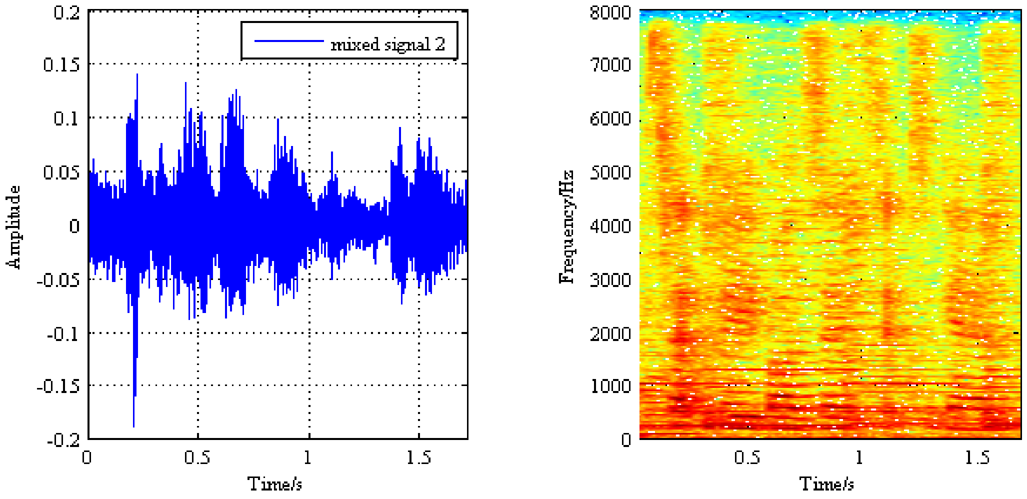

Figure 5. Time evolution and spectrograms of mixed signal $m_2(n)$

In experiment, BSS step size $\eta = 0.2$ is used and for volterra adaption, we used step size $\mu_0 = \mu_1 = \mu_2 = 0.05$, $L = 64$.

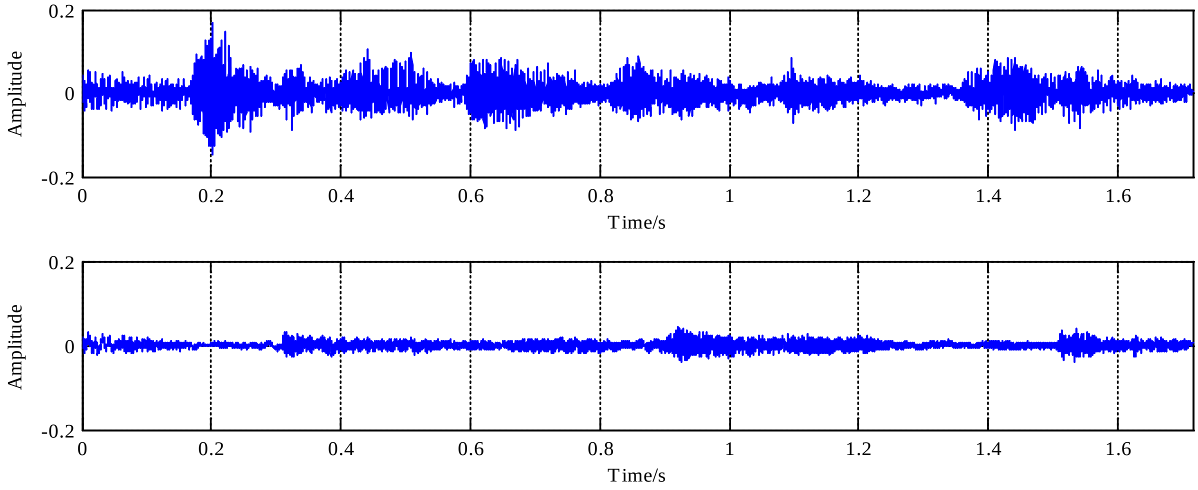

Figure 6. Time evolution of the proposed framework (top: before the proposed framework, bottom: after the proposed framework)

In Figs. 6, we show the time evolution of the output speech signal obtained by the proposed framework: in the top of figure, we show the waveform before AEC, and in the bottom we show the result after AEC process. Its ERLE is shown in Figs. 7.

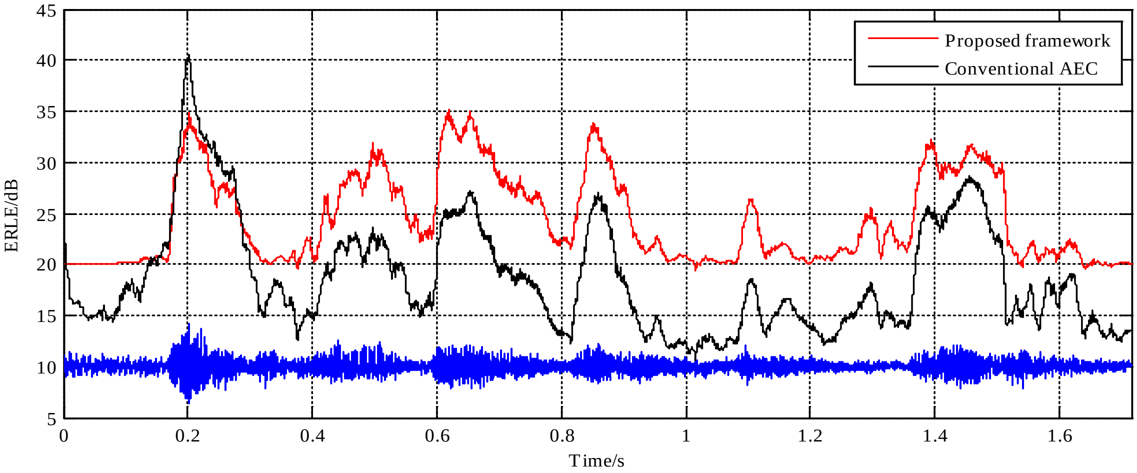

Figure 7. ERLE criterion evaluation of the proposed framework and conventional AEC

According to the obtained results of the ERLE criterion in Fig. 7, the conventional AEC produces a comparable result with the proposed framework at the beginning of signal. With the convergence of adaptive algorithm, it is clear that the proposed framework has the higher ERLE values in comparison with the conventional AEC.

5. Conclusion

In this paper, the BSS-AEC framework has been proposed to cancel the echo in double-talk scenario without the DTD. The DTD is displaced by BSS method, which is used to separate near-end speech in double talk scenario. To validate the proposed framework, we have carried out intensive experiments based on objective criteria ERLE. The simulation results have confirmed the superiority of the proposed framework in term of the used criteria.

References

[1] Sakai Y, Akhtar M T. The acoustic echo cancelation using blind source separation to reduce double-talk interference [C]//2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC). IEEE, 2014: 323 - 326.

[2] Gunther J. Learning echo paths during continuous double-talk using semi-blind source separation [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 20 (2): 646 - 660.

[3] Yang J M, Sakai H. A new adaptive filter algorithm for system identification using independent component analysis [J]. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, 2007, 90 (8): 1549 - 1554.

[4] Nesta F, Matassoni M. Blind source extraction for robust speech recognition in multisource noisy environments [J]. Computer Speech & Language, 2013, 27 (3): 703 - 725.

[5] Kumar M N, PrasannaVani V, Saravanan S. Evaluation of LMS and volterra adaptive filters for AEC[C]//2017 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES). IEEE, 2017: 1 - 6.

[6] Stenger A, Rabenstein R. Adaptive Volterra filters for nonlinear acoustic echo cancellation [C]//NSIP. 1999: 679 - 683.