BLIND SOURCE SEPARATION AND ACOUSTIC ECHO CANCELLATION: A UNIFIED FRAMEWORK

Muhammad Z. Ikram

Systems and Applications R&D Center, Texas Instruments, Inc., Dallas, TX 75243
mzi@ti.com

ABSTRACT

We explore interesting connections between blind source separation (BSS) and acoustic echo cancellation (AEC), and develop a framework where the AEC problem is transformed and solved as a BSS problem. We show that by careful selection of the BSS algorithm the double-talk (DT) problem in AEC is solved without the need to use a DT detector or a step-size controller. Furthermore, the echo cancellation performance is maintained even during single-talk when only the far-end speaker is active. The algorithm converges to the true echo path much faster than the normalized least-mean squares adaptation. Moreover, the proposed algorithm does not require a knowledge of the echo-tail length and is robust against under estimation of the echo-filter length. The simple implementation and fast convergence of the proposed method makes it a suitable candidate for implementation on low-power general purpose DSPs.

Index Terms— Blind signal separation, Acoustic Echo Cancellation, Double Talk, Impulse Response

1. INTRODUCTION

Acoustic echo cancellation (AEC) is a classical problem with a history that dates back to 1960s [1]. Since then a number of attractive methods have been proposed to address various aspects of this problem. One of the issues that has remained at the forefront of AEC development is its handling of the double talk (DT). DT refers to simultaneous activation of the far-end and the near-end speakers. The adaptive filter in an echo canceler is designed to cancel only the far-end echo, and any presence of near-end signal strongly influences its convergence. The resulting mis-convergence causes the far-end listener to hear its own echo at the onset of the near-end speech, which is annoying and undesirable. Ever since the development of the first echo canceller several methods have been proposed to tackle the DT. Certain solutions handle it by restricting the communication to only one way; i.e., half-duplex. In other cases, a DT detector is used that freezes filter adaptation in the presence of DT [1]. While half-duplex communication is not desirable in many situations the DT detector based AEC has its own issues as well. First, in order to benefit from frozen adaptation, the DT detector has to be efficient in its estimation of start and end of the near-end speech. Any mis-detection may lead to near-end speech leakage in the uplink communication. Secondly, the echo cancellation may suffer if the echo path changes during the time of frozen adaptation.

In this paper we address the DT problem in AEC from a practical point of view. We develop an interesting relationship between AEC and blind source separation (BSS) and propose to transform the AEC problem into a BSS problem. BSS methods, in general, aim to separate spatially distributed speech signals using multiple microphones [2]. In other applications they help to separate desired speech embedded in noise. In this work, we will show how a two-microphone BSS setup can be used to solve AEC even when only one microphone is available. We prove that by treating the near-end speaker and the far-end speaker as two statistically uncorrelated speech sources, application of BSS leads to their separation and subsequent recovery of the near-end speech when DT is present. Furthermore, we formulate the problem such that the algorithm convergence is not disturbed when the conversation switches from single talk to DT or vice versa. In other words, the algorithm requires no step-size control or a DT detector for convergence control. We will also show that the algorithm is robust against under determination of the echo-filter length. The fast convergence of the algorithm and its simple processing makes it a good fit for handheld and hands-free speech communication applications.

2. AEC AND THE DOUBLE-TALK PROBLEM

Fig. 1 shows an AEC setup, where $s_1(n)$ and $s_2(n)$ are the far-end and near-end speakers, respectively. The echo path from the loudspeaker to the microphone is modeled by a length-L FIR filter $\mathbf{h}_{21}(n) = \left[h_{21}^0(n), h_{21}^1(n), \ \ldots, h_{21}^{L-1}(n)\right]^T$, where the superscript on the filter coefficient denotes the tap index and $[\cdot]^T$ denotes transposition. Likewise, the adaptive filter of length P is denoted by $\mathbf{w}_{21}(n) = \left[w_{21}^0(n), w_{21}^1(n), \ldots, w_{21}^{P-1}(n)\right]^T$. The reason for using the subscript "21" will become clear as we describe the BSS problem in Section 3.

The update of the filter $\mathbf{w}_{21}(n)$ using the well-known nor-

malized least-mean squares (NLMS) algorithm is governed by [1]

$$w_{21}^{k}(n) = w_{21}^{k}(n-1) + \alpha \frac{e(n)x_1^*(n-k)}{\sum_{i=0}^{P-1}|x_1(n-i)|^2}, \quad (1)$$

for $k = 0, \ldots, P-1$, where α is the adaptation constant and $(\cdot)^*$ denotes complex conjugation. The echo is cancelled when $\mathbf{w}_{21}(n) = \mathbf{h}_{21}(n)$. In the event of the DT there is a tendency for the AEC to diverge resulting in leakage of the far-end echo in the uplink channel. Handling of the DT in AEC has always remained an area of interest within the speech research community. Several notable methods have been proposed for this purpose. In [3], [4], [5], and [6] an extra component is employed in the AEC to help avoid filter divergence during DT. This component could be in the form of a DT detector or a step-size controller. In [7] the authors presented a variable step-size NLMS (VSS-NLMS) algorithm that is robust against DT. The automatic step-size control mechanism of this method halts the adaptation during instances of DT. Though the algorithm did not require explicit DT detection, its convergence rate was slow and of the order of the NLMS method.

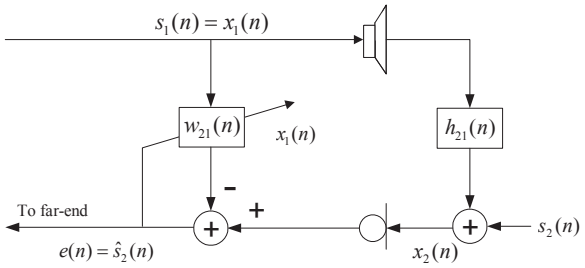

Fig. 1. AEC setup.

3. AEC IN BSS FRAMEWORK

BSS relies on the statistical independence of the source signals to separate them from their mixtures. A 2-input 2-output BSS setup is shown in Fig. 2, where the signals $s_1(n)$ and $s_2(n)$ are mixed using FIR filters $\mathbf{h}_{ij}(n)$, $i, j = 1, 2$. The two mixed outputs $x_1(n)$ and $x_2(n)$ are recorded by two spatially separated microphones. In an acoustic environment, the two signals corresponds to two speakers or to a speaker and a noise source, whereas the four FIR filters model the room impulse responses from each source to a microphone location. The objective of BSS is to separate and recover the two source signals using only the two recorded microphone outputs. The separation is carried out using a symmetrical un-mixing stage where the FIR filters $\mathbf{w}_{ij}(n)$, $i, j = 1, 2$ are adapted such that the two outputs $\widehat{s}_1(n)$ and $\widehat{s}_2(n)$ are statistically independent. Mathematically, this statistical independence is represented by [2]

$$E\left\{\widehat{s}_1(n)\widehat{s}_2^*(n-m)\right\} = 0; \ \forall m, \quad (2)$$

where $E\{\cdot\}$ denotes statistical expectation and $\widehat{(\cdot)}$ denotes an estimate. At convergence, the two mixed signals $s_1(n)$ and $s_2(n)$ are recovered at the output.

With the BSS setup of Fig. 2 in mind if we recall the DT problem in AEC we realize that, in fact, we try to separate a mixture of the near-end and far-end speech signals using the microphone output $x_2(n)$. Since the near-end and far-end signals originate from different speakers it is safe to assume that they are statistically independent. We can, therefore, use BSS to recover the near-end speech signal from its mixture with the far-end echo. The setup of Fig. 2 can be simplified to a functionally-equivalent AEC setup of Fig. 3 by choosing $\mathbf{h}_{11} = 1$, $\mathbf{h}_{22} = 1$, and $\mathbf{h}_{12} = 0$. Consequently, the un-mixing section will have $\mathbf{w}_{11} = 1$, $\mathbf{w}_{22} = 1$, and $\mathbf{w}_{12} = 0$. Note that the resulting BSS setup shown in Fig. 3 is equivalent to Fig. 1. We note that one of the microphone inputs is replaced by the reference signal $x_1(n) = s_1(n)$ whereas the other microphone input is the sum of the echo signal and the near-end speech. The filter $\mathbf{w}_{12}(n)$ is adapted such that the signals $s_1(n)$ and $s_2(n)$ are statistically independent and uncorrelated. An adaptation method will be presented in Section 4.

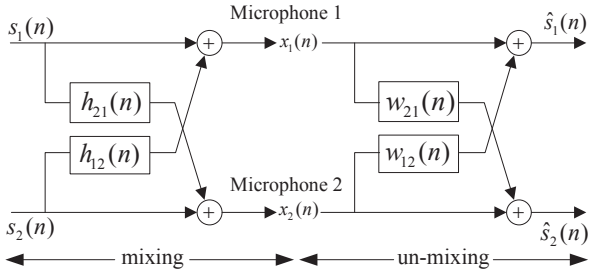

Fig. 2. A two-channel BSS setup.

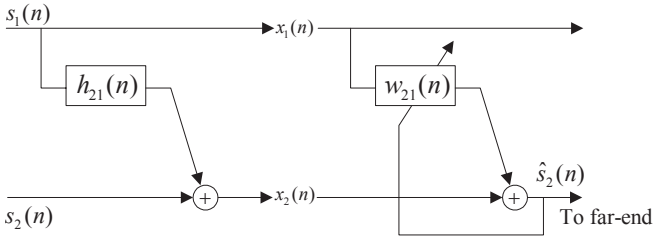

Fig. 3. A functionally equivalent AEC setup.

In [8] the author proposed an approach somewhat similar to ours in using BSS for handling the DT. However, the method in [8] works only during the continuous DT period; i.e., when both the far-end as well as the near-end speech is present. Furthermore, slow convergence rate was reported in [8] and the author suggested to use the method as a complement to a conventional and fast-converging AEC method operating during single talk.

4. THE DT-ROBUST AEC ALGORITHM

Let us stack the two speech signals in the vector form as $\mathbf{s}(n) = [s_1(n), s_2(n)]^T$. The two received signals are modeled as convolutive mixtures of the two speech signals and can be expressed in matrix-vector form as

$$\mathbf{x}(n) = \mathbf{H}(n) * \mathbf{s}(n), \tag{3}$$

where $\mathbf{x}(n) = [x_1(n), x_2(n)]^T$ is the received signal vector, $\mathbf{H}(n) = \begin{bmatrix} 1 & 0 \\ h_{21}(n) & 1 \end{bmatrix}$ is the (2×2) mixing filter matrix of L-point impulse responses, and $*$ is the convolution operator. In general, the aim of the BSS method is to find a (2×2) un-mixing filter $\mathbf{W}(n)$ of length P that separates the two sources

$$\widehat{\mathbf{s}}(n) = \mathbf{W}(n) * \mathbf{x}(n), \tag{4}$$

where $\mathbf{W}(n) = \begin{bmatrix} 1 & 0 \\ w_{21}(n) & 1 \end{bmatrix}$. Our objective in the equivalent AEC setup of Fig. 3 is to extract only the source signal $s_2(n)$.

We propose to transform the time-domain convolutive mixture $\mathbf{x}(n)$ in (3) to an instantaneous mixture in the frequency domain by computing its T-point short-time Fourier transform

$$\mathbf{x}(\omega, m) = \mathbf{H}(\omega)\mathbf{s}(\omega, m), \tag{5}$$

where m is the block index. In practice, $\mathbf{x}(\omega, m)$ can be obtained as follows

$$\mathbf{x}(\omega, m) = \sum_{\tau=0}^{T-1} \gamma(\tau)\mathbf{x}(\beta T m + \tau)e^{-j2\pi\omega\tau/T}, \tag{6}$$

for $\omega = 1, \ldots, T$, where $\gamma(\tau)$ is a window function and β $(0 < \beta \leq 1)$ is the data overlap factor. The covariance matrix $\mathbf{R}_{\mathbf{x}}(\omega, k)$, assuming ergodicity of the received data, can be estimated using M, possibly overlapping, blocks of $\mathbf{x}(\omega, m)$ as follows

$$\widehat{\mathbf{R}}_{\mathbf{x}}(\omega, k) = \frac{1}{M}\sum_{m=0}^{M-1} \mathbf{x}(\omega, Mk + m)\mathbf{x}^H(\omega, Mk + m), \tag{7}$$

for $k = 0, \ldots, K-1$, where $(\cdot)^H$ denotes conjugate transposition. The un-mixing filter matrix $\mathbf{W}(\omega)$ decorrelates the estimated source signals $\widehat{s}_1(n)$ and $\widehat{s}_2(n)$ by diagonalizing their covariance matrix given by

$$\mathbf{\Lambda}_{\widehat{\mathbf{s}}}(\omega, k) = \mathbf{W}(\omega)\widehat{\mathbf{R}}_{\mathbf{x}}(\omega, k)\mathbf{W}^H(\omega). \tag{8}$$

We showed in [2] that for non-stationary signals, we can write independent decorrelation equations (8) for K sufficiently separated time intervals. The un-mixing filter $\mathbf{W}(\omega)$ for each frequency bin ω $(\omega = 1, \ldots, T)$ that simultaneously satisfies the K decorrelation equations can then be obtained using an over-determined least-squares solution

$$\widehat{\mathbf{W}}(\omega) = \arg\min_{\mathbf{W}(\omega)} \sum_{k=1}^{K} \|\mathbf{V}(\omega, k)\|^2, \tag{9}$$

where $\|\cdot\|^2$ is the squared Frobenius norm (sum of squares of all elements) and the error

$$\begin{aligned} \mathbf{V}(\omega, k) &= \mathbf{W}(\omega)\widehat{\mathbf{R}}_{\mathbf{x}}(\omega, k)\mathbf{W}^H(\omega) \\ &\quad -\mathrm{diag}\left[\mathbf{W}(\omega)\widehat{\mathbf{R}}_{\mathbf{x}}(\omega, k)\mathbf{W}^H(\omega)\right], \end{aligned} \tag{10}$$

where $\mathrm{diag}[\cdot]$ is the diagonal matrix formed by extracting the diagonal elements of the matrix argument. Note that in (8) there are only two unknowns: $W_{21}(\omega)$ and $\Lambda_{s_2}(\omega, k)$, whereas, because of the symmetry of the matrix, there are three constraints in the same equation. This significantly simplifies the BSS problem and makes is extremely suitable for real-time implementation since K can be chosen as small as 1 to satisfy the over-determined condition.

In practice, the least-squares solution to (9) can be obtained using the well-known steepest descent algorithm

$$\begin{aligned} \mathbf{W}^{(l+1)}(\omega) &= \mathbf{W}^{(l)}(\omega) - \mu(\omega) \\ &\quad \cdot \frac{\partial}{\partial \mathbf{W}^{(l)^*}(\omega)} \left\{ \sum_{k=1}^{K} \|\mathbf{V}^{(l)}(\omega, k)\|^2 \right\} \end{aligned} \tag{11}$$

for $\omega = 1, \ldots, T$. Following [2], we use a step size of the form

$$\mu(\omega) = \frac{\eta}{\sum_{k=1}^{K} \|\widehat{\mathbf{R}}_{\mathbf{x}}(\omega, k)\|^2}, \tag{12}$$

where η is a normalized step size. At each iteration, we only update one of the off-diagonal elements, $W_{21}(\omega)$, of $\mathbf{W}(\omega)$ while retaining the remaining elements at their initial values. As we shall see in the next section, this key step ensures that the filter convergence remain on track even if the near-end speech is not present in the single-talk case. We will also see that the independent step size at each frequency promotes overall convergence of the algorithm. At convergence, time-domain signal $s_2(n)$ is obtained using inverse Fourier transform.

5. SIMULATION RESULTS

In our experiments, we will focus on the following aspects of the AEC: (1) performance in the presence of DT, (2) convergence rate, and (3) performance when the adaptive-filter length is under estimated. A male speaker is at the far end of the conversation whereas a female speaker is at the near end. Fig. 4 shows the two speech signals. The echo length is about 30msec. The average far-end to DT ratio is about 5dB. It is seen that the near-end speech is active only for time duration from 10sec to about 21sec. This helps us to evaluate the AEC performance going from single talk to the DT mode and back. We will compare the proposed method against the NLMS-based AEC, which is a widely used algorithm employed in most commercial applications.

We will use adaptive-filter misalignment as the performance indicator. It is defined as

$$20\log 10\frac{\|\mathbf{h}_{21} - \mathbf{w}_{21}\|}{\|\mathbf{h}_{21}\|}. \tag{13}$$

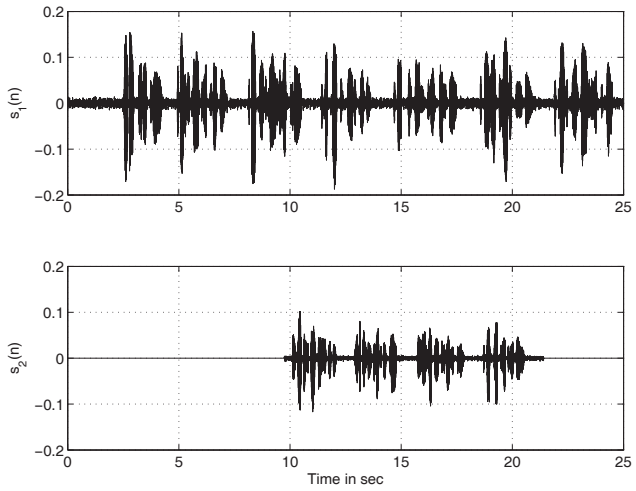

Fig. 4. The two speech signals used in the experiment. $s_1(n)$ is the far-end speech, whereas $s_2(n)$ is the near-end speech.

For NLMS adaptation, we used $\alpha = 0.5$, whereas for the proposed method we used $\eta = 0.5$, $T = 512$, $\beta = 0.5$ (50% overlap), and $M = 10$. A Hamming window function was used for $\gamma(\tau)$ and we began our experiments with a value of $P = 256$.

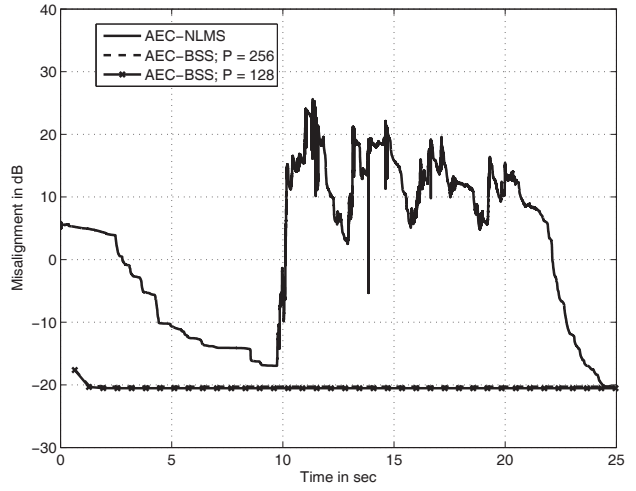

Fig. 5. Misalignment plots of the NLMS-based AEC and the proposed method ($P = 256$ and $P = 128$)

.

Fig. 5 shows the divergence of the NLMS-based AEC at the onset of the near-end speech. The algorithm then re-tries to converge as soon as the near-end speaker pauses during conversation. On the other hand, the BSS-based AEC converges much faster than the NLMS method and stays converged during and after the DT. We also note that the misalignment remains the same even when the adaptive filter length P is reduced to 128 taps.

6. CONCLUSION

We transformed the AEC problem to a BSS framework and solved it using a frequency-domain second-order statistics based method. We showed that by minimizing the statistical correlation between the far-end and the near-end speech signals in BSS guarantees that the echo is cancelled during single talk as well as during DT. No extra control or component is required to deal with the DT. The variable step-size assignment in frequency domain promotes convergence and the resulting AEC algorithm is robust against under determination of the adaptive filter-length. All these properties makes this algorithm an excellent candidate for implementation on low-power DSPs.

7. REFERENCES

[1] S. Gay and J. Benesty, Eds., *Acoustic Signal Processing for Telecommunication*, Norwell, MA: Kluwer Academic, 2000.

[2] M. Z. Ikram and D. R. Morgan, "Permutation inconsistency in blind speech separation: Investigation and solutions," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 1, pp. 1–13, Jan. 2005.

[3] K.-H. Lee, J.-H. Chang, N. S. Kim, S. Kang, and Y. Kim, "Frequency-domain double-talk detection based on the Gaussian mixture model," *IEEE Signal Processing Letters*, vol. 17, no. 5, pp. 453–456, May 2010.

[4] R. Oka, K. Fujii, and M. Muneyasu, "A step size control method steadily reducing acoustic echo even during double talk," in *Proc. Int. Symp. Intelligent Signal Processing and Communication Systems*, 2008.

[5] H. Buchner, J. Benesty, T. Gänsler, and W. Kellermann, "Robust extended multidelay filter and double-talk detector for acoustic echo cancellation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1633–1643, Sept. 2006.

[6] T. Gänsler, S. L. Gay, M. M. Sondhi, and J. Benesty, "Double-talk robust fast converging algorithms for network echo cancellation," *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 6, pp. 656–663, Nov. 2000.

[7] C. Paleologu, S. Ciochină, and J. Benesty, "Double-talk robust VSS-NLMS algorithm for under-modeling acoustic echo cancellation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2008, pp. 245–248.

[8] J. Gunther, "Learning echo paths during continuous double-talk using semi-blind source separation," To Appear in *IEEE Trans. Audio, Speech, and Language Processing*, 2011.