OPTIMAL JOINT LINEAR ACOUSTIC ECHO CANCELATION AND BLIND SOURCE SEPARATION IN THE PRESENCE OF LOUDSPEAKER NONLINEARITY

Mehrez Souden[1,*] *and Zicheng Liu*[2]
[1] INRS-ÉMT, 800, de la Gauchetière Ouest, Suite 6900, Montréal, H5A 1K6, Qc, Canada.
[2] Microsoft Research, One Microsoft Way, Redmond WA 98052.

ABSTRACT

Acoustic echoes represent a major source of discomfort in hands free, full-duplex, communication systems. The problem becomes particularly difficult when the loudspeakers are nonlinear as considered in this paper. In contrast to the single-microphone linear and nonlinear acoustic echo cancelation techniques, we take advantage of the spatial diversity offered by the microphone arrays. Indeed, having a set of microphones and multiple sources (i.e., the near and far ends) that can be active at the same time, this problem can be solved using a blind source separation (BSS) algorithm. The performance of the BSS can be further improved when combined with a linear acoustic echo canceler (LAEC). In this paper, we study the potentials of joint BSS and LAEC to cancel the echo signals in two schemes. In the first scheme, the BSS is deployed as a front-end and has a twofold function: reducing the acoustic echo and creating a linearly transformed echo reference that is used by the LAEC as a post-processor. In the second scheme, the BSS operates on multiple LAECs outputs to further reduce the residual echo from the target signal. We show that the first scheme outperforms the second one.

Index Terms— Acoustic echo cancelation, nonlinear loudspeaker, blind source separation, microphone arrays.

1. INTRODUCTION

Hands free and full duplex communication systems require a set of microphones and loudspeakers to be deployed within the same enclosure, thereby leading to a coupling between these two types of devices. This fact represents a major source of discomfort for the users. Hence, the design of an efficient acoustic echo canceler is extremely important.

The problem of acoustic echo cancelation is classical yet still very challenging due to the hostile nature of the acoustic environment. Indeed, major hindrances to the development of a reliable acoustic echo canceler include, but are not limited to: excessive duration of the acoustic path in reverberant enclosures [1], long double talk periods, and loudspeaker imperfections that translate into unknown transformations of the echo-line signals [2, 3]. The latter is attributed to space and cost constraints resulting in small-sized and low-quality (cheap) loudspeakers [2, 3]. Classically, the problem of acoustic echo cancelation is handled through single channel linear processing [1]. To deal with loudspeaker nonlinearity, earlier efforts have been dedicated to model and modify classical single-channel LAEC accordingly [2, 3, 4]. Nowadays, microphone arrays are becoming commonplace in communication devices and they have been shown to be of great advantage in several applications including source localization, noise reduction, and blind source separation [1]. Therefore, one can expect them to help in acoustic echo cancelation too. A notable earlier effort that took advantage of multiple microphones was presented in [5] where Herbordt et al. showed the advantage of using beamforming jointly with LAEC. However, beamforming requires an accurate voice (near-end) activity detector in addition to the target source location to prevent its cancelation. In contrast, the BSS represents a good alternative since it does not require such prior knowledge [6, 7].

In this paper, we study the potentials of joint BSS and LAEC in echo cancelation. In particular, we are interested in the case where the loudspeakers are nonlinear since this is still a challenging problem and we believe that the BSS can play a significant role to solve it. Indeed, due to the coexistence of both near and far ends, the problem of acoustic echo cancelation is seen as the well known *cocktail party* that can be addressed by BSS. To take advantage of the known far-end signal, one would expect that the LAEC can further improve the performance. Intuitively, there are two schemes combining BSS and LAEC. In the first scheme, a BSS is deployed as a front-end and has a twofold function: reducing the acoustic echo and creating a linearly transformed echo reference that is used by the LAEC (deployed as a post-processor). In the second scheme, a LAEC is applied to each channel and the BSS is deployed as a back-end to operate on the LAEC outputs and reduce the residual echo. As far as we know, there hasn't been any reported study on this problem. One related work was done by Low and Nordholm [8] who proposed a post-processor for BSS to jointly suppress echo and interference. However, echo suppression increases the target signal distortion and we would rather use a LAEC. Moreover, neither the loudspeaker nonlinearity issue nor the second scheme were considered therein. The purpose of this paper is to study the two schemes of joint BSS and LAEC in the general case of arbitrary loudspeaker nonlinearity. Our study shows that the first scheme outperforms the second one and that the BSS is very effective in echo cancelation especially in the presence of loudspeaker nonlinearity.

2. PROBLEM STATEMENT AND ASSUMPTIONS

We consider the case where a talker (near-end), generating a signal $s(t)$, and a loudspeaker, playing a signal $z(t)$, are located within the same enclosure in addition to a set of N microphones. The microphones outputs are then given by

$$\mathbf{x}(t) = \mathbf{g} * s(t) + \mathbf{h} * z(t) + \mathbf{v}(t) \quad (1)$$

where $\mathbf{x}(t) = [x_1(t) \ \cdots x_N(t)]^T$, $\mathbf{g} = [g_1 \ \cdots g_N]^T$ and $\mathbf{h} = [h_1 \ \cdots h_N]^T$ are the propagation paths of $s(t)$ and $z(t)$ toward the microphones, and $\mathbf{v}(t) = [v_1(t) \ \cdots v_N(t)]^T$ represent some additive noise components. $*$ is the convolution operator. Our objective is to recover a clean version of $s(t)$. It has to be pointed out that we consider the general case where the far-end signal denoted as $e(t)$ is *arbitrarily transformed* by the loudspeaker. In other words, we suppose that $z(t) = f[e(t)]$ where $f[\cdot]$ is a nonlinear function. This is typically the case of small-sized or cheap loudspeakers [2, 3, 4].

*The first author performed the work while at Microsoft Research.

3. LINEAR ACOUSTIC ECHO CANCELATION AND EFFECT OF LOUDSPEAKER NONLINEARITY

Traditionally, acoustic echo cancelation techniques consist in imitating the echo path during the echo-only periods to create a copy of this signal as seen by the microphone and subtracting it from the overall microphone output before transmitting it to the far-end. A robust echo path estimator is, thus, required. To this end, several single channel estimation techniques have been proposed so far (see [1] and references therein). The main difference between most of them is attributed to the choice of the adaptation step-size in the classical LMS algorithm [1, 9]. To understand the effect of the loudspeaker nonlinearity, we focus on the basic LMS algorithm (for a LAEC deployed after the nth microphone, $n = 1 \cdots N$)

$$\hat{\mathbf{h}}_n(t+1) = \hat{\mathbf{h}}_n(t) + \mu_n \varepsilon_n(t)\mathbf{e}(t), \tag{2}$$

where $\varepsilon_n(t) = x_n(t) - \hat{\mathbf{h}}_n^T(t)\mathbf{e}(t)$, $\mathbf{e}(t) = [e(t) \ \cdots e(t-L+1)]^T$, $\hat{\mathbf{h}}_n(t) = [\hat{h}_n^{(0)}(t) \ \cdots \hat{h}_n^{(L-1)}]^T$, and L is the number of taps in the echo path $\mathbf{h}_n$ whose estimate is $\hat{\mathbf{h}}_n(t)$ at instant t. Assuming that the function $f[\cdot]$ can be decomposed into two terms $f[e(t)] = e(t) + g[e(t)]$, where $g[e(t)]$ is a pure nonlinear transform of $e(t)$ and ignoring the presence of the additive noise, the error signal can be written as[1]

$$\varepsilon_n(t) = \left[\mathbf{h}_n - \hat{\mathbf{h}}_n(t)\right]^T \mathbf{e}(t) + \mathbf{h}_n^T g[\mathbf{e}(t)]. \tag{3}$$

By also defining $\mathbf{c}_n(t) = \mathbf{h}_n(t) - \hat{\mathbf{h}}_n(t)$ and using (2), we obtain

$$\mathbf{c}_n(t+1) = \mathbf{c}_n(t) - \mu_n \mathbf{e}(t)\mathbf{e}^T(t)\mathbf{c}_n(t) - \mu_n \mathbf{e}(t) g\left[\mathbf{e}^T(t)\right]\mathbf{h}_n. \tag{4}$$

If we assume that $\mathbf{c}_n(t)$ is independent of $\mathbf{e}(t)$ as in [9], we obtain

$$\mathrm{E}\{\mathbf{c}_n(t+1)\} = (\mathbf{I} - \mu_n \mathbf{R}_{ee})\,\mathrm{E}\{\mathbf{c}_n(t)\} - \mu_n \mathbf{R}_{eg}\mathbf{h}_n, \tag{5}$$

where $\mathbf{R}_{ee} = \mathrm{E}\{\mathbf{e}(t)\mathbf{e}^T(t)\}$ and $\mathbf{R}_{eg} = \mathrm{E}\{\mathbf{e}(t) g[\mathbf{e}^T(t)]\}$. It can be easily shown from (5) that $\mathrm{E}\{\mathbf{c}_n(t)\} = (\mathbf{I} - \mu_n \mathbf{R}_{ee})^t \mathrm{E}\{\mathbf{c}_n(0)\} - \mu_n \sum_{k=0}^{t} (\mathbf{I} - \mu_n \mathbf{R}_{ee})^k \mathbf{R}_{eg}\mathbf{h}_n$. A common practice is to choose $0 < \mu_n < 1/\gamma_{\max}$ with $\gamma_{\max}$ being the largest eigenvalue of $\mathbf{R}_{ee}$ to ensure the convergence of the LMS [9]. Unfortunately, it is clear that with such choice $\mathrm{E}\{\mathbf{c}_n(t)\} \sim \mu_n \sum_{k=0}^{t} (\mathbf{I} - \mu_n \mathbf{R}_{ee})^k \mathbf{R}_{eg}\mathbf{h}_n \neq \mathbf{0}$ when $t \to \infty$ and $g[e(t)]$ is correlated with $e(t)$, meaning that the channel estimate in this case is biased. One can also easily verify that the nonlinear term results in increased mean square error. In [2, 3], nonlinear Volterra filters were used to obtain a more reliable estimate. In [4], a raised cosine function was used to compensate the loudspeaker nonlinearity. Here, we use the BSS to take advantage of the spatial dimension since multiple microphones are deployed.

A reliable double talk detector (DTD) is also extremely important in a LAEC. Indeed, the channel estimate has to be performed only during the absence of the near-end. The most common and robust DTD is given by [1]

$$\zeta_n = \sqrt{\frac{\mathbf{h}_n^T \mathbf{R}_{ee}\mathbf{h}_n}{\mathrm{E}\{x_n^2(t)\}}}. \tag{6}$$

In practice, however, $\mathbf{h}_n$ is not available and one would rather replace it with $\hat{\mathbf{h}}_n(t)$, hoping that $\hat{\mathbf{h}}_n(t)$ has converged enough. When ζ_n is larger than a certain threshold, we decide that the near-end is absent and update the channel estimate. The choice of this threshold is ad-hoc. A major limitation that can be immediately seen when considering (6) in the case of nonlinear loudspeaker is that $\hat{\mathbf{h}}_n(t)$ is a biased estimator. Clearly, the loudspeaker nonlinearity also affects the DTD. The resulting channel estimate during double talk periods may even diverge due to erroneous decisions.

[1] We ignore the tail effect and assume a memoryless nonlinearity [2, 3].

4. CONVOLUTIVE BLIND SOURCE SEPARATION

When many sources are active at the same time, a natural solution to separate them consists in using a BSS technique. To alleviate this complexity, one of the commonly used approaches is to decompose the received signals into frequency bins and processing each of them separately. This apparent simplifying transformation is not without caveats. Indeed, processing each frequency bin independently from the others leads to well known permutation and scaling issues that have to be solved in order to avoid speech distortions (due to scaling indetermination) and frequency swapping (due to permutation). A good survey on BSS is provided in [6]. Without loss of generality, we choose the non-stationarity-based BSS algorithm that was proposed in [7]. Herein, we briefly describe this algorithm. First, (1) is transformed to the frequency domain via the $\mathcal{T}$-length short time Fourier transform (STFT)

$$\underline{\mathbf{x}}(\omega,\tau) = \underline{\mathbf{A}}(\omega)\underline{\mathbf{y}}(\omega,\tau) + \underline{\mathbf{v}}(\omega,\tau), \tag{7}$$

where $\underline{\mathbf{A}}(\omega) = \left[\underline{\mathbf{g}}(\omega)\ \underline{\mathbf{h}}(\omega)\right]$, $\underline{\mathbf{y}}(\omega,\tau) = [\underline{s}(\omega,\tau)\ \underline{z}(\omega,\tau)]^T$. $\underline{\mathbf{x}}(\omega,\tau)$, $\underline{\mathbf{g}}(\omega)$, $\underline{\mathbf{h}}(\omega)$, $\underline{s}(\omega,\tau)$, $\underline{z}(\omega,\tau)$, and $\underline{\mathbf{v}}(\omega,\tau)$ are the STFT's of $\mathbf{x}(t)$, $\mathbf{g}$, $\mathbf{h}$, $s(t)$, $z(t)$, and $\mathbf{v}(t)$, respectively. ω stands for the discrete frequency in $\{1, ..., \mathcal{T}\}$, and τ is the frame index. The separation of $\underline{s}(\omega,\tau)$ and $\underline{z}(\omega,\tau)$ is achieved by applying the separation matrix $\underline{\mathbf{W}}(\omega)$ to $\underline{\mathbf{x}}(\omega,\tau)$

$$\hat{\underline{\mathbf{y}}}(\omega,\tau) = \underline{\mathbf{W}}(\omega)\underline{\mathbf{x}}(\omega,\tau). \tag{8}$$

The separation algorithm proposed in [7] exploits the non-stationarity of the speech signals. In fact, $\underline{\mathbf{W}}(\omega)$ can be found by jointly diagonalizing the the cross-power spectrum (PSD) matrices of the output data estimated at K time intervals. In our work, the kth ($k \in \{1, \cdots, K\}$) estimate of the PSD matrix of the observations, $\hat{\underline{\mathbf{R}}}_{xx}(\omega,k)$, is calculated using the modified Welch's periodogram [10] instead of using a simple time averaging for better accuracy. Then, $\underline{\mathbf{W}}(\omega)$ is given by [7]

$$\underline{\mathbf{W}}(\omega) = \arg\min_{\tilde{\underline{\mathbf{W}}}(\omega)} \sum_{k=1}^{K} \left\|\underline{\mathbf{E}}(\omega,k)\right\|_F^2, \tag{9}$$

where $\underline{\mathbf{E}}(\omega,k) = \mathrm{offdiag}\left[\tilde{\underline{\mathbf{W}}}(\omega)\hat{\underline{\mathbf{R}}}_{xx}(\omega,k)\tilde{\underline{\mathbf{W}}}^H(\omega)\right]$. (9) can only be solved in an iterative way using the gradient descent with a step-size $\mu(\omega)$ that has to be properly chosen [7]

$$\underline{\mathbf{W}}^{(l+1)}(\omega) = \underline{\mathbf{W}}^{(l)}(\omega) - \mu(\omega)\sum_{k=1}^{K} \underline{\mathbf{E}}(\omega,k)\underline{\mathbf{W}}^{(l)}(\omega)\hat{\underline{\mathbf{R}}}_{xx}(\omega,k). \tag{10}$$

To solve the scaling and permutation issues, the estimate of the separation matrix at the lth iteration is modified as

$$\begin{aligned} \left[\underline{\mathbf{W}}^{(l)}(\omega)\right]_{pp} &= 1, \\ \left[\underline{\mathbf{W}}^{(l)}(\omega)\right]_{pq} &\leftarrow \mathbf{F}\mathbf{T}\mathbf{F}^{-1}\left[\underline{\mathbf{W}}^{(l)}(\omega)\right]_{pq}, \end{aligned} \tag{11}$$

where $\mathbf{F}$ is the $\mathcal{T} \times \mathcal{T}$ DFT matrix, $\mathbf{T}$ is the $\mathcal{T} \times \mathcal{T}$ truncation matrix with $[\mathbf{T}]_{i_1 i_2} = 1$ if $i_1 = i_2$ and $i_1 \leq Q$ and $[\mathbf{T}]_{i_1 i_2} = 0$ otherwise, where $Q < \mathcal{T}$ is the length of filter. This algorithm may converge to a local minimum since the optimization criterion is not convex [6]. However, we have empirically found that with well distributed microphones and sources (e.g., having one microphone in the vicinity of each speaker as in teleconferencing rooms [11]), this algorithm leads to well separated signals.

5. JOINT BLIND SOURCE SEPARATION AND LINEAR ACOUSTIC ECHO CANCELATION

Due to presence of both near and far ends, the problem of acoustic echo cancelation can be seen as the well known *cocktail party*. Hence, one can naturally think about using BSS to separate both signals. The performance of the BSS can be further improved when a LAEC is deployed as a post-processor. This scheme will be termed "BSS-AEC ". Conversely, one can also imagine the scenario where each microphone is equipped with a LAEC. The echo residual can still be separated from the near-end signal using BSS. This scheme will be termed "AEC-BSS."

5.1. First Scheme: BSS-AEC

The principle of this scheme is depicted in Fig. 1. A BSS algorithm is deployed as a first stage to process the microphone outputs and yield $y_1(t)$ and $y_2(t)$. A global permutation can occur even after solving the frequency swapping problem using (11). Indeed, one cannot immediately distinguish between the estimate of the echo and the one corresponding to the near-end. To deal with this issue, we measure the similarity between $e(t)$ and each of the BSS outputs. Statistically, this similarity can be measured using the second order statistics (the normalized cross-correlation or the cross-coherence) or the higher order statistics (cross-cumulants). We empirically found that regardless of the loudspeaker nonlinearity, we are able to distinguish between both output signals by simply using the cross-coherence (the block "Coh" in Fig. 1). After solving the global permutation, we notice that due to the imperfections of the first stage (e.g., statistics estimation errors, unsolved permutations for some frequency bins, and effect of the noise), $y_1(t)$ and $y_2(t)$ are expressed as

$$y_1(t) = \left[\sum_{n=1}^{N} w_{1n} * g_n\right] * s(t) + \underbrace{\left[\sum_{n=1}^{N} w_{1n} * h_n\right] * z(t)}_{z_r(t)} \quad (12)$$

$$y_2(t) = \left[\sum_{n=1}^{N} w_{2n} * h_n\right] * z(t) + \underbrace{\left[\sum_{n=1}^{N} w_{2n} * g_n\right] * s(t)}_{s_r(t)} \quad (13)$$

where $w_{mn};\ m = 1,\ 2$ and $n = 1, \cdots, N$, are the entries of the separation matrix, $z_r(t)$ is the residual echo in the separated near-end signal and $s_r(t)$ is the residual near-end. Ideally, we would like to have $z_r(t) = s_r(t) = 0$. Unfortunately, this is never the case in practice. It is very important to note that $z_r(t)$ is a linearly transformed version $z(t)$ which is unknown in practice. If we further neglect $s_r(t)$, we see that $z_r(t)$ and $y_2(t)$ are linearly transformed copies of each others. Clearly, the BSS creates a reference signal, $y_2(t)$, for the residual echo that can be removed using a LAEC. The advantage of using $y_2(t)$ as a reference to further reduce the echo is that, theoretically, the loudspeaker nonlinearity has no effect. This comes at the price of some distortions of the final output near-end signal due to the residual $s_r(t)$ in $y_2(t)$. However, as long as the BSS outputs are well separated, this distortion remains negligible. Note that one can still use the echo-line signal as depicted in Fig. 1. Essentially, this would be beneficial when the loudspeaker nonlinearity is low. In our case, we decide to use $e(t)$ as a reference if the coherence between $e(t)$ and $y_2(t)$ is higher than an empirical threshold $Th = 0.8$. Otherwise, we use $y_2(t)$. To sum up, the BSS preprocessing has two advantages in this scheme: it reduces the inter-signals interference and creates a reference signal, $y_2(t)$, to be used by the LAEC.

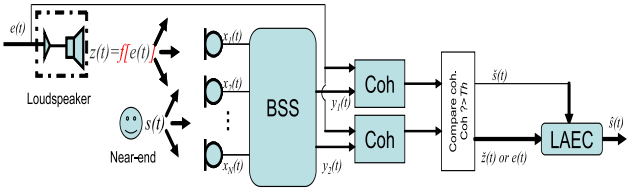

Fig. 1. A sketch of the BSS-AEC.

5.2. Second Scheme: AEC-BSS

This scheme is depicted in Fig. 2. Each microphone is equipped with a LAEC. In Section 3, we have shown how the loudspeaker nonlinearity leads to a biased estimate of the echo path and affects the DTD. However, by deploying a BSS algorithm as a second stage, one hopes to further reduce the residual echo. Actually, the BSS is designed to create two signals which are as independent as possible. The first one is dominated by the near-end while the second is dominated by the residual echo. We measure the similarity (using the coherence, block "Coh" in Fig. 2) between each of the BSS outputs and $e(t)$ to blindly distinguish between both signals. Let us ignore the additive noise and consider the decomposition of $z(t)$ into a linear and nonlinear components as in Section 3. The output of the nth LAEC can be expressed as

$$\varepsilon_n(t) = g_n * s(t) + h_n * g\left[e(t)\right] + \left(h_n - \hat{h}_n\right) * e(t). \quad (14)$$

Unfortunately, the behavior of the LAEC cannot be predicted due to the nonlinearity that is seen as an additive noise which can be arbitrarily correlated to $e(t)$. In the extreme case where $g\left[e(t)\right]$ is uncorrelated with $e(t)$, the loudspeaker nonlinearity still has a negative impact on the LAEC and consequently on the BSS algorithm because there are essentially three uncorrelated sources. In the general case, no straightforward conclusion can be drawn to predict the performance of the BSS as a post-processor. Numerical evaluations show that the performance of the AEC-BSS is entirely dependent on the LAEC stage.

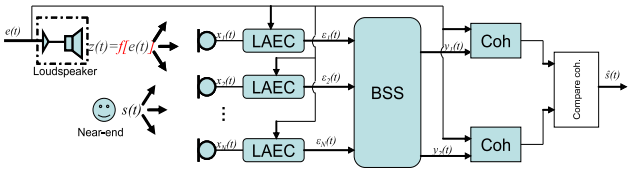

Fig. 2. A sketch of the AEC-BSS.

6. NUMERICAL EXAMPLES

We evaluate the performance of both schemes when the IPNLMS [1] is used as a LAEC in addition to the non-stationarity-based BSS described in Section 4. A reverberant room with dimensions: length $= 3.5$ m, width $= 4.6$ m, and height $= 3.6$ m ($x \times y \times z$) is simulated using the image method [12] with $T_{60} \approx 250$ ms. The near-end is around 15 seconds-long female speech located at $(1.5, 1.5, 1)$ m and the far-end signal is a male speech with a loudspeaker located at $(2.5, 1.5, 1)$ m. A computer generated white Gaussian noise is also added such that the signal to noise ratio is 30 dB. Two microphones are deployed at the locations $(1.5, 2, 1)$ m and $(2.5, 2, 1)$ m. Note that this configuration is typical in a teleconferencing room with distributed microphones (see [11] and references therein). The signals

are sampled at 8 kHz and cut into 75% overlapping frames of length 4096 samples each. The lengths of all filters were set to 2048 taps. We tested several loudspeaker non-linearities including the following particular function $f[s(t)] = (1-\lambda)s(t)+(1-\lambda)\tanh[4s(t)]$. The choice of the "tanh" function is motivated by the fact that it models very well the soft-clipping of the speech signals (large energy components are saturated while the low energy ones are almost unchanged) [4]. In order to evaluate the echo attenuation, we use the echo return loss enhancement (ERLE) [2, 3, 4]. We also measure the distortion of the estimated (echo-free) near-end in terms of the log-likelihood ratio (LLR) between the estimated near-end and its noise-free version captured by the nearest microphone [13]. This signal distortion measure has been shown to be very correlated to the human subjective evaluation (with a correlation factor of around 0.61) [13]. We compare the performance of both schemes in addition to the BSS only and the standard single channel-LAEC when microphone 1 and microphone 2 are used (i.e., AEC-1 and AEC-2, respectively). Note that the results are smoothed enough by using large averaging windows as can be understood from Figs. 3 and 4.

In the first simulation setup, we set $\lambda = 1$ (i.e., perfectly linear loudspeaker). In Fig. 3, we clearly see that AEC-2 provides the highest ERLE values after convergence and during the "echo only" period. This is understood because the microphone 2 is closer to the loudspeaker. The performance of AEC-2 remarkably deteriorates during the double talk period. This is due to the false detections which are more likely to happen with microphone 2 while the performance of the AEC-1 are almost steady. The BSS alone provides the poorest ERLE. The AEC-BSS performance is comparable to the AEC-1. Interestingly, the BSS-AEC outperforms all algorithms during the double talk period. The LLR measure shows that the BSS-AEC leads to very low signal distortions in contrast to the AEC-2. This fact is explained by the reverberant path between the near-end and the second microphone.

In the second simulation setup, we set $\lambda = 0.4$. In Fig. 4, we see that the BSS and BSS-AEC are not sensitive to the loudspeaker nonlinearity. Conversely, the ERLE achieved by all other methods is dramatically degraded. Indeed, the arbitrary nonlinearity caused by the loudspeaker severely deteriorates the performance of the LAEC even during the "echo only" periods. During the double talk period, the estimate of the echo path is not reliable and leads to the malfunctioning of the double talk detector. The ERLE achieved by AEC-1 and AEC-2 is reduced and the deployment of the BSS as a post-processor does not improve their performance. In all cases, we conclude that the BSS-AEC performs the best.

7. CONCLUSIONS

In this paper, we investigated the capabilities of the BSS for acoustic echo cancelation. Our study took into consideration the loudspeaker nonlinearity. Specifically, we proposed two schemes. The first one consists in blindly separating the sources then using a LAEC to further reduce the echo. The second one consists in equipping each microphone with a LAEC and postprocessing the outputs of these filters by a BSS algorithm. From our experiments, we found that the first scheme is more effective for acoustic echo cancelation.

8. REFERENCES

[1] Y. Huang, J. Benesty, and J. Chen, *Acoustic MIMO signal processing*. Springer-Verlag, Berlin, Germany, 2006.

[2] F. Kuech and W. Kellermann, "Partitioned block-brequency domain adaptive second order Volterra filter," *Trans. Signal Process.*, vol. 53, no. 2, pp. 564-575, Feb. 2005.

[3] A. Guerin, G. Faucon, and R. Le Bouquin-Jeannes, "Nonlinear acoustic echo cancellation based on Volterra filters," *IEEE Trans. Speech and Audio Processing*, vol. 11, no. 6, pp. 672-683, Nov. 2003.

[4] H. Dai and W.P. Zhu, "Compensation of loudspeaker nonlinearity in acoustic echo cancelation using raised cosine function," *Trans. Circuits and Systems-II*, vol. 53, no. 11, pp. 1190-1194, Nov. 2003.

[5] W. Herbordt, H. Buchner, W. Kellermann, "An acoustic human-machine front-end for multimedia applications," *European Journal on Applied Signal Process.*, vol. 2003, no. 1, pp. 1-11, Jan. 2003.

[6] M.S. Pedersen, J. Larsen, U. Kjems, and L.C. Parra, "A survey of convolutive blind source separation methods," in *Multichannel Speech Processing Handbook*, Eds., J. Benesty, M.M. Sondhi, and A. Huang, Springer-Verlag, Berlin, 2007.

[7] L. Parra and C. Spence, "Convolutive blind source separation of non-stationary sources," *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 3, pp. 320-327, May 2000.

[8] S.Y. Low and S. Nordholm, "A blind approach to joint echo and noise cancellation," in *Proc. IEEE ICASSP*, Mar. 2005, vol. 3, pp. 69-72.

[9] P.S.R. Diniz, *Adaptive filtering: algorithms and practical implementation,* Kluwer Academic Publishers, Norwell, MA, USA, 2002.

[10] A.V. Oppenheim and R.W. Schafer, *Discrete-time signal processing*, Upper Saddle River, NJ: Prentice-Hall, 1999.

[11] J.P. Dmochowski, Z. Liu, and P.A. Chou," Blind source separation in a distributed microphone meeting environment for improved teleconferencing," in *Proc. IEEE ICASSP*, Mar. 2008, pp. 89-92.

[12] J.B. Allen and D.A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal Acoust. Society of America*, vol. 65, no. 4, pp. 943-950, Apr. 1979.

[13] Y. Hu and P.C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

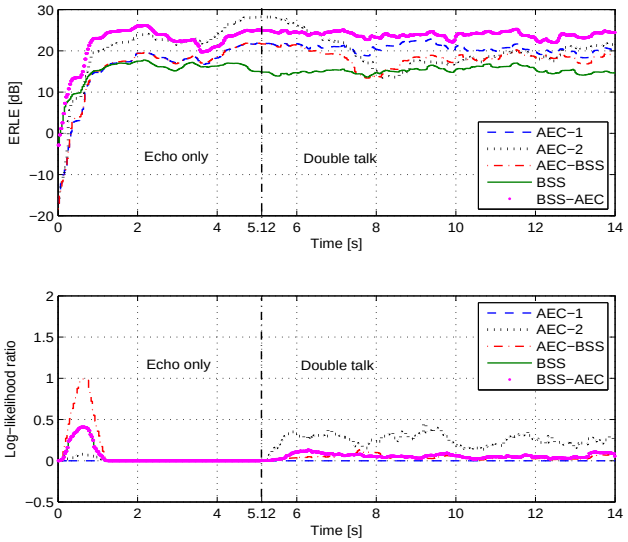

Fig. 3. Top: ERLE vs. time, Bottom: LLR vs. time; $f[s(t)] = s(t)$.

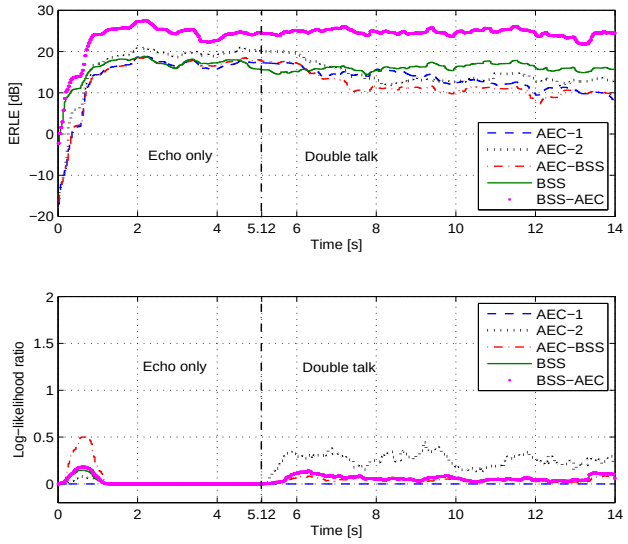

Fig. 4. Top: ERLE vs. time, Bottom: LLR vs. time; $f[s(t)] = 0.6s(t) + 0.4\tanh[4s(t)]$.