USE OF DECORRELATION PROCEDURE FOR SOURCE AND ECHO SUPPRESSION

Ted S. Wada[†]*, Shigeki Miyabe*[‡]*, Biing-Hwang (Fred) Juang*[†]

[†] Center for Signal and Image Processing, Georgia Institute of Technology, 75 Fifth Street NW, Atlanta, GA, 30332, USA

[‡] Nara Institute of Science and Technology, 630-0192, Takayama-cho 8916-5, Ikoma-shi, Nara, Japan

e-mail: [†]{twada,juang}@ece.gatech.edu, [‡]shige-m@is.naist.jp

ABSTRACT

The so-called *semi-blind* source separation (SBSS) is an extension of BSS when some signals can be directly and individually measured. SBSS can be used to implement multi-channel acoustic cancellation (MCAEC) without the need for double-talk detection. However, traditional MCAEC approaches based on least mean-square (LMS) lead to a non-unique solution due to correlated sources formed when a single far-end signal is captured by multiple microphones and played out through the near-end loudspeakers. The same non-uniqueness problem can be also expected during SBSS. In this study, the effect of using a decorrelation procedure for mitigating the non-uniqueness problem by applying a nonlinearity or a noise to the far-end reference signals is measured when BSS and stereophonic AEC (SAEC) are simultaneously implemented through SBSS. The simulation results show that the benefit, in terms of the misalignment and the echo return loss enhancement (ERLE), of using a decorrelation procedure can be significant for SBSS while not interfering with the source separation performance at all.

Index Terms— semi-blind source separation, multi-channel acoustic echo cancellation, non-uniqueness problem

1. INTRODUCTION

Blind source separation (BSS) has become a very popular and effective method for canceling interfering signals from unknown sources when there are multiple input channels. Most of the latest BSS techniques are based on the independent component analysis (ICA) [1] that achieves separation of mixed signals by maximizing the statistical independence between output signals. Recently, there have been several attempts at combining BSS with acoustic echo cancellation (AEC) in the frequency-domain [2–4]. Such a combination of BSS and AEC is appropriately referred to as semi-BSS (SBSS) since the signals from the far-end sources that are to be played back through near-end loudspeakers are known prior to mixing at the near-end, thus *semi-blind*, and can be used directly for adaptation of the unmixing filter. The clear advantage of implementing AEC through BSS techniques is that the double-talk detection (DTD) is no longer needed since no distinction is made between local and remote signal sources, whereas the disadvantage is that the current BSS algorithms are much more computationally expensive than the traditional AEC algorithms.

Just as in BSS that deals with separation of multiple sources, we are also interested in the multi-channel AEC (MCAEC) for the identification of multiple echo paths and the cancellation of signals that go through them. It is well known that MCAEC algorithms based on the least mean-square (LMS) theory suffer from the non-uniqueness problem caused by the ill-conditioning of the autocorrelation matrix formed from multiple reference signals that are correlated [5–7]. The problem is especially significant when only one remote source is active, and it makes the tracking of changes in the echo paths much more difficult as the LMS-based adaptive algorithms may not be able to converge to a new non-unique solution fast enough. Thus to alleviate the ill-conditioning of the matrix, the reference signals can be pre-processed by applying some multi-channel form of nonlinearity to decrease the correlation between them [6, 7]. This process, however, can degrade the quality of playback of the received signals, and it incurs a cap on how much decorrelation can be reasonably achieved without adding perceptually noticeable distortion. Another straight-forward method is to add white noise to the signals, the procedure of which is also limited by the degradation in the signal-to-noise ratio (SNR). Many other decorrelation methods are discussed in [7].

In this paper, we explore the effect of applying a decorrelation procedure when BSS and stereophonic AEC (SAEC) are implemented together through SBSS. Although the ICA-based BSS algorithms are inherently more robust to ill-conditioning than the traditional LMS-based algorithms, as ICA allows the optimization of not only the second order but also the higher order statistics of output signals, we should expect that they are affected in some way by highly correlated signal sources during the system identification process. Also, it is not clear how much side effect the decorrelation procedure has on the source separation performance.

The paper is organized as follows. First, the semi-blind source separation scheme is described that achieves both source separation and echo cancellation. Second, the SBSS and the traditional AEC algorithms to be tested and methods for evaluating the effects of decorrelation procedures are presented. Finally, simulation results and discussions are given, followed by concluding remarks at the end.

2. SEMI-BLIND SOURCE SEPARATION

Figure 1 shows the overall configuration for a combination of stereophonic teleconferencing and SBSS. We assume here that all of the sources are stationary in order to eliminate the problem of tracking the room response change due to moving sources. Then the time-invariant instantaneous mixing in the frequency-domain that occurs at the far-end and at the near-end can be summarized together as

$$\begin{bmatrix} \boldsymbol{X}_n(f,t) \\ \boldsymbol{X}_f(f,t) \end{bmatrix} = \begin{bmatrix} \boldsymbol{H}_n(f) & \boldsymbol{H}_f(f) \\ 0 & \boldsymbol{G}(f) \end{bmatrix} \begin{bmatrix} \boldsymbol{S}_n(f,t) \\ \boldsymbol{S}_f(f,t) \end{bmatrix}, \quad (1)$$

where $\boldsymbol{S}_n(f,t)$ and $\boldsymbol{S}_f(f,t)$ are the short-time Fourier transform (STFT) of the near and the far-end sources $\boldsymbol{s}_n(t) = [\boldsymbol{s}_1(t), \boldsymbol{s}_2(t)]^T$

[†] Supported in part by the National Science Foundation Award IIS-0534221. [‡] Supported by the JSPS Fellowship for the Young Scientists.

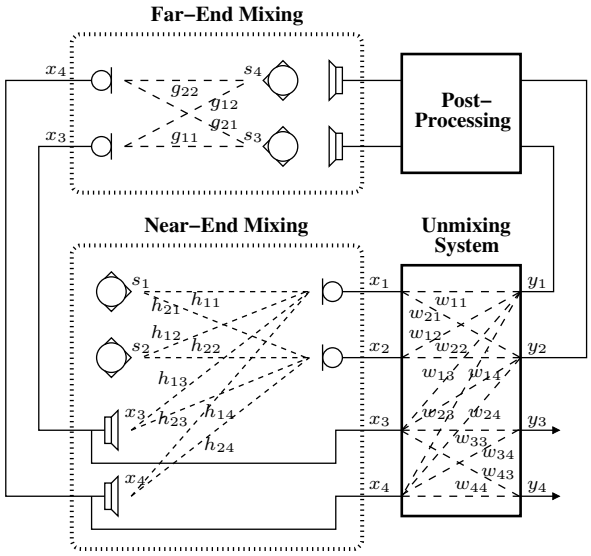

Fig. 1. Configuration for stereophonic teleconferencing and semi-blind source separation (SBSS).

and $\boldsymbol{s}_f(t) = [\boldsymbol{s}_3(t), \boldsymbol{s}_4(t)]^T$, respectively; $\boldsymbol{X}_n(f,t)$ and $\boldsymbol{X}_f(f,t)$ are the STFT of the locally mixed signals and the reference signals $\boldsymbol{x}_n(t) = [\boldsymbol{x}_1(t), \boldsymbol{x}_2(t)]^T$ and $\boldsymbol{x}_f(t) = [\boldsymbol{x}_3(t), \boldsymbol{x}_4(t)]^T$, respectively; and $\boldsymbol{H}_n(f)$, $\boldsymbol{H}_f(f)$, $\boldsymbol{G}(f)$ are the STFT of the near and the far-end room impulse responses $\boldsymbol{h}_n(t) = [[\boldsymbol{h}_{11}(t), \boldsymbol{h}_{21}(t)], [\boldsymbol{h}_{21}(t), \boldsymbol{h}_{22}(t)]]^T$, $\boldsymbol{h}_f(t) = [[\boldsymbol{h}_{13}(t), \boldsymbol{h}_{23}(t)], [\boldsymbol{h}_{14}(t), \boldsymbol{h}_{24}(t)]]^T$, and $\boldsymbol{g}(t) = [[\boldsymbol{g}_{11}(t), \boldsymbol{g}_{21}(t)], [\boldsymbol{g}_{21}(t), \boldsymbol{g}_{22}(t)]]^T$, respectively. The 4-by-4 unmixing matrix $\boldsymbol{W}(f)$ must be estimated such that

$$\begin{bmatrix} \boldsymbol{Y}_n(f,t) \\ \boldsymbol{Y}_f(f,t) \end{bmatrix} = \begin{bmatrix} \boldsymbol{W}_1(f) & \boldsymbol{W}_2(f) \\ 0 & \boldsymbol{W}_3(f) \end{bmatrix} \begin{bmatrix} \boldsymbol{X}_n(f,t) \\ \boldsymbol{X}_f(f,t) \end{bmatrix}. \quad (2)$$

The optimal $\boldsymbol{W}(f)$ can be obtained by employing ICA and maximizing the statistical independence between all of the components in the output vector $\boldsymbol{y}(t) = [\boldsymbol{y}_n(t), \boldsymbol{y}_f(t)]^T = [\boldsymbol{y}_1(t), \boldsymbol{y}_2(t), \boldsymbol{y}_3(t), \boldsymbol{y}_4(t)]^T$ as demonstrated in [2], [3], and [4].

The matrix $\boldsymbol{W}_1(f) = [[w_{12}(f), w_{21}(f)], [w_{21}(f), w_{22}(f)]]^T$ is for source separation while $\boldsymbol{W}_2(f) = [[w_{13}(f), w_{23}(f)], [w_{14}(f), w_{24}(f)]]^T$ is for echo cancellation, which by design is performed *after* the source separation. On the other hand, the exact form of $\boldsymbol{W}_3(f) = [[w_{33}(f), w_{43}(f)], [w_{34}(f), w_{44}(f)]]^T$ is arbitrary, as it can be constrained to be an identity matrix [2], a diagonal matrix, or a full-matrix [4]. Although the output $\boldsymbol{y}_f(t) = [\boldsymbol{y}_3(t), \boldsymbol{y}_4(t)]^T$ is not necessary for the stereo teleconferencing purpose, it is still used during ICA-based optimization. In this study, we will use a diagonal matrix as motivated by [3] that did not assume $\boldsymbol{y}_f(t) = \boldsymbol{x}_f(t)$. We feel that a full matrix will counteract the decorrelation procedure since in such a case (2) implies that $\boldsymbol{y}_f(t)$ is a linear combination of $\boldsymbol{x}_f(t)$. We also believe that the use of a more general diagonally constrained matrix instead of the identity matrix is better suited for simultaneous optimization of the source separation and the echo cancellation performances through ICA.

In the traditional MCAEC framework, changes in the far-end room response $\boldsymbol{G}(f)$ can also disrupt the filter adaptation that is performed at the near-end (i.e. when an active source suddenly switches from one to another at the far-end) [5–7]. For such a situation, the adaptation of $\boldsymbol{W}_3(f)$ may be designed accordingly to track the changes in $\boldsymbol{G}(f)$. However, such a topic is beyond the scope of this paper and is not addressed at this time.

3. EVALUATION METHODS

In this study, the SBSS algorithm in [3] is extended to include not just one but two reference channels. The residual echo signal *without* the source separation can be obtained by taking the matrix product

$$\boldsymbol{E}(f,t) = [\boldsymbol{I}, \boldsymbol{W}_1^{-1}(f)\boldsymbol{W}_2(f)][\boldsymbol{X}_n^T(f,t), \boldsymbol{X}_f^T(f,t)]^T, \quad (3)$$

where $\boldsymbol{I}$ is a 2-by-2 identity matrix. The pre-multiplication of $\boldsymbol{W}_2(f)$ by $\boldsymbol{W}_1^{-1}(f)$ follows the projection back procedure [3, 8], which further enhances the source separation performance and is a part of the post-processing indicated in Fig. 1. The algorithm from WinEC presented in [9] is used to implement the traditional SAEC. WinEC uses a coherence-based DTD along with an outlier-robust adaptive algorithm that limits the effect of double-talk leakage before the filter adaptation is suspended. The residual echo suppression is not used for comparison purpose with SBSS.

For both SBSS and WinEC, the following "half-wave rectifying" decorrelation nonlinearity is used [7]:

$$x_3'(t) = x_3(t) + \alpha\frac{x_3(t) + |x_3(t)|}{2}, \quad (4)$$

$$x_4'(t) = x_4(t) + \alpha\frac{x_4(t) - |x_4(t)|}{2}. \quad (5)$$

The amount of decorrelation, thus the degree of signal distortion, is controlled by the factor α, which can be as large as 0.5 before the stereo perception is affected [7]. A simple procedure of inserting the additive white Gaussian noise (AWGN) to the reference channels is also tested as another decorrelation method.

Three types of performance measures are used. First, the *true* echo return loss enhancement (tERLE), which measures the mean square error (MSE) performance, for the i^{th} near-end microphone channel (i = 1,2) is defined as

$$\text{tERLE}(i) = 10\log_{10}\frac{\|\boldsymbol{x}_i(t) - \sum_{k=1}^{2}\boldsymbol{h}_{ik}^T(t)\boldsymbol{s}_k(t)\|^2}{\|\boldsymbol{e}_i(t) - \sum_{k=1}^{2}\boldsymbol{h}_{ik}^T(t)\boldsymbol{s}_k(t)\|^2} \text{ (dB)}, \quad (6)$$

which is simply the traditional ERLE calculated after removing the near-end speech so that the true reduction in the echo can also be calculated during the double-talk. Next, the misalignment, which measures the system identification performance, for the ij^{th} echo path (i = 1,2, and j = 3,4) is defined as

$$\text{Misalignment}(i,j) = 10\log_{10}\frac{\|\boldsymbol{h}_{ij}(t) - \hat{\boldsymbol{h}}_{ij}(t)\|^2}{\|\boldsymbol{h}_{ij}(t)\|^2} \text{ (dB)}, \quad (7)$$

where the impulse response estimate $\hat{\boldsymbol{h}}_{ij}(t)$ for SBSS can be obtained by taking the inverse Fourier transform of $-\boldsymbol{W}_1^{-1}(f)\boldsymbol{W}_2(f)$. Finally, the signal-to-interference ratio (SIR), which measures the source separation performance, for the i^{th} local source (i = 1, 2, 3, 4) is defined as

$$\text{SIR}_{\text{in}}(i) = 10\log_{10}\frac{\sum_{k=1}^{2}(\boldsymbol{h}_{ki}^T(t)\boldsymbol{s}_i'(t))^2}{\sum_{k=1}^{2}(\sum_{j\neq i}\boldsymbol{h}_{kj}^T(t)\boldsymbol{s}_j'(t))^2} \text{ (dB)}, \quad (8)$$

$$\text{SIR}_{\text{out}}(i) = 10\log_{10}\frac{\sum_{k=1}^{2}(\boldsymbol{r}_{ki}^T(t)\boldsymbol{s}_i'(t))^2}{\sum_{k=1}^{2}(\sum_{j\neq i}\boldsymbol{r}_{kj}^T(t)\boldsymbol{s}_j'(t))^2} \text{ (dB)}, \quad (9)$$

$$\text{SIR}(i) = \text{SIR}_{\text{out}}(i) - \text{SIR}_{\text{in}}(i) \text{ (dB)}, \quad (10)$$

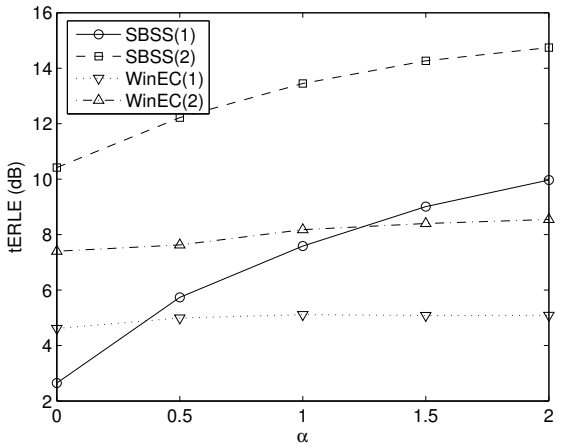

Fig. 2. Average tERLE when used with half-wave rectifying decorrelation nonlinearity while varying the distortion factor α for one (1) or two (2) far-end talkers.

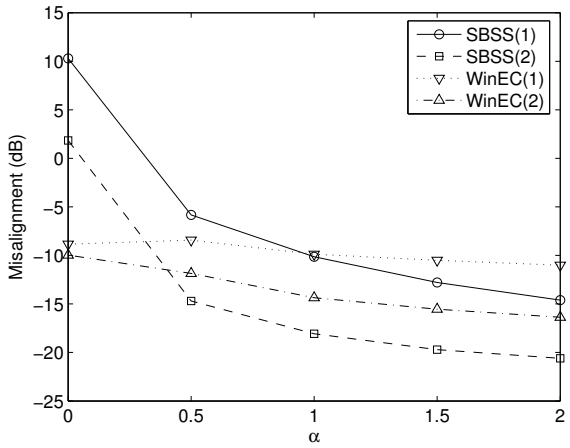

Fig. 4. Average misalignment when used with half-wave rectifying decorrelation nonlinearity while varying the distortion factor α for one (1) or two (2) far-end talkers.

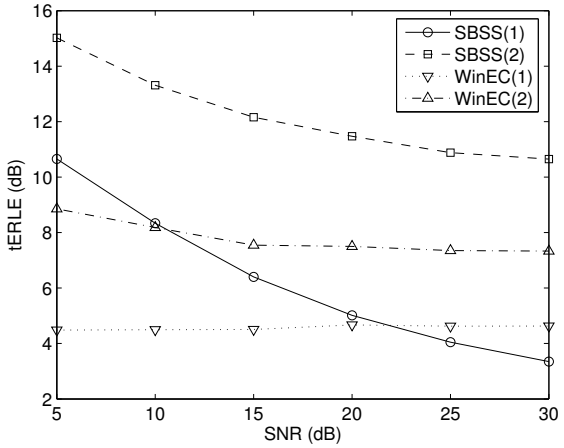

Fig. 3. Average tERLE when used with additive white Gaussian noise while varying SNR for one (1) or two (2) far-end talkers.

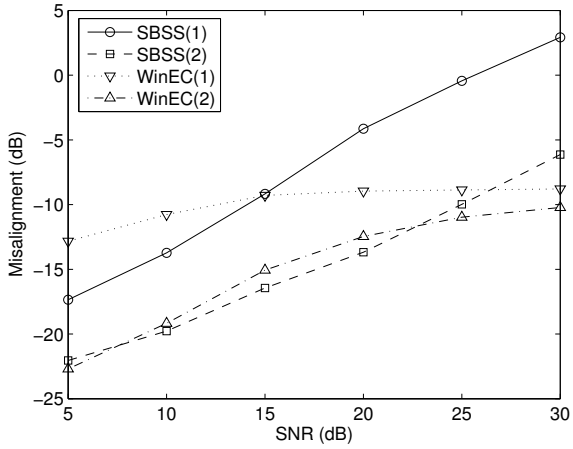

Fig. 5. Average misalignment when used with additive white Gaussian noise while varying SNR for one (1) or two (2) far-end talkers.

where $\boldsymbol{s}'_i(t) = \boldsymbol{s}_i(t)$ for i = 1,2, or $\boldsymbol{s}'_i(t) = \boldsymbol{x}_i(t)$ for i = 3,4, and $\boldsymbol{r}_{ki}(t)$ is an overall impulse response for the k^{th} output that encompasses the near-end room response, the unmixing filter response, and the projection back filter response.

4. SIMULATION RESULTS

Two sets of impulse responses, one for far-end and another for near-end, were measured in a controlled acoustic environment and used for simulation. The impulse responses were truncated to 128 ms in length. A recorded background sound containing air conditioner noise was scaled to 30 dB SNR and added to the near-end microphone mixtures to simulate a more realistic acoustic condition. 16 kHz TIMIT recordings were used to simulate either one or two far-end talkers and always two near-end talkers. The individual TIMIT speeches were concatenated with silence added between them such that there were roughly 25% overlap of speeches at the far-end, 25% overlap between the far-end and the near-end signals (i.e. during double-talk), and 75% overlap of speeches at the near-end. Single-talk (i.e. when only the far-end sources are active) was enforced for the first two seconds in order for WinEC to converge sufficiently. SBSS was ran for 500 iterations in a batch mode (i.e. off-line) for 10 seconds of simulated signals. WinEC was adapted frame-by-frame (i.e. on-line) for also 10 seconds. The performance measures were averaged over ten simulations using different sets of speeches and across all possible microphone channels and echo paths. The tERLE was calculated only during voice activity for both SBSS and WinEC.

It must be noted before any comparisons are made between SBSS and WinEC that since we are not dealing with the problem of tracking the changes in the room responses, the advantage of using the decorrelation procedure is mostly lost for WinEC. Also, it is unfair to compare SBSS directly against WinEC since the number of iterations used by the SBSS's adaptive algorithm is not limited. WinEC is used here mainly to show what is possible with the

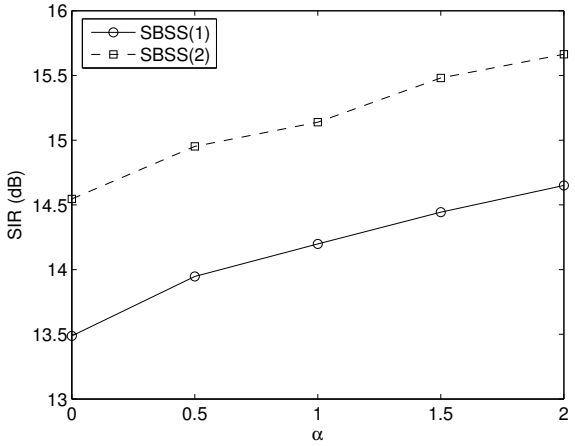

Fig. 6. Average SIR when used with half-wave rectifying decorrelation nonlinearity while varying the distortion factor α for one (1) or two (2) far-end talkers.

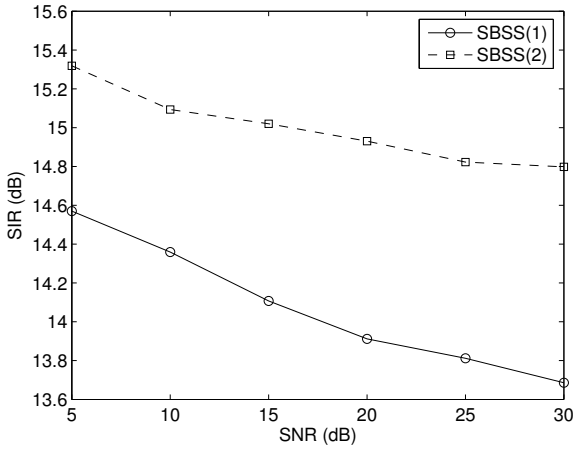

Fig. 7. Average SIR when used with additive white Gaussian noise while varying SNR for one (1) or two (2) far-end talkers.

traditional real-time SAEC approach.

The average tERLE plots for SBSS and WinEC as a function of α and SNR are shown in Fig. 2 and 3. The number in parentheses (1 or 2) next to algorithm names in the figures correspond to the number of far-end talkers. We see from the figures that as the degree of decorrelation is increased (i.e. as α is increased or SNR is decreased), WinEC shows only a small improvement in the tERLE, whereas SBSS exhibits between 3 to 8 dB gain in the tERLE. Also, having two far-end sources helps both SBSS and WinEC achieve better echo cancellation, which is consistent with the idea that multiple active far-end sources that are separated spatially should help decorrelate the reference signals.

The average misalignment plots for SBSS and WinEC as a function of α and SNR are shown in Fig. 4 and 5. The figures show that the improvement due to the decorrelation procedures is much more significant for SBSS-based system identification than for the traditional LMS-based approach. The reduction in the misalignment can be over 20 dB for SBSS compared to 10 dB for WinEC. Again, the two-sources case produces lower misalignment than the one-source case for both SBSS and WinEC.

Finally, the average SIR for SBSS as a function of α and SNR is plotted in Fig. 6 and 7. Although not as dramatic as the improvement in the ERLE and the misalignment, some gain in the SIR is observed for both decorrelation methods. The overall improvement in the SIR is consistent with the notion that better echo cancellation should also help the adaptation of unmixing filter for source separation by reducing the correlation between the output signals caused by the residual echo.

5. CONCLUSION

The effect of using a decorrelation procedure in order to alleviate the non-uniqueness problem during semi-blind source separation (SBSS) for the purpose of stereophonic acoustic echo cancellation (SAEC) was examined. Simulated experiment showed that either applying the half-wave rectifying decorrelation nonlinearity or adding the white Gaussian noise to the reference signals can decrease the misalignment by over 20 dB for SBSS. The corresponding improvement of between 3 to 8 dB in the true echo return loss enhancement (tERLE), which is the traditional ERLE calculated by removing the near-end speech during the double-talk, was shown. Small yet observable improvement in the source separation performance, measured in terms of the signal-to-interference ratio (SIR), was also achieved after the inclusion of the decorrelation procedures.

6. REFERENCES

[1] A. Hyvärinen et al., *Independent Component Analysis*, Wiley Interscience, 2001.

[2] M. Joho et al., "Combined blind/nonblind source separation based on the natural gradient," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 8, pp. 236–238, 2001.

[3] S. Miyabe et al., "Barge-in and noise-free spoken dialogue interface based on sound field control and semi-blind source separation," in *Proc. EUSIPCO*, 2007, pp. 232–236.

[4] J. Even et al., "Frequency-domain semi-blind signal separation: application to the rejection of internal noises," in *Proc. ICASSP*, 2008, pp. 157–160.

[5] M.M. Sondhi et al., "Stereophonic acoustic echo cancellation - an overview of the fundamental problem," *IEEE Signal Proc. Lett.*, vol. 2, no. 15, pp. 148–151, 1995.

[6] J. Benesty et al., "A better understanding and an improved solution to the specific problems of stereophonic acoustic echo cancellation," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 2, pp. 156–165, 1998.

[7] J. Benesty et al., "Adaptive algorithms for MIMO acoustic echo cancellation," in *Audio signal processing for next-generation multimedia communication systems*, Y. Huang and J. Benesty, Eds., pp. 119–147. Kluwer Academic, 2004.

[8] N. Murata et al., "An on-line algorithm for blind source separation on speech signals," in *Proc. NOLTA*, 1998, vol. 3, pp. 923–926.

[9] T. Gänsler et al., "The WinEC: a real-time hands-free stereo communication system," in *Audio signal processing for next-generation multimedia communication systems*, Y. Huang and J. Benesty, Eds., pp. 171–193. Kluwer Academic, 2004.