A NEW ALGORITHM FOR JOINT BLIND SIGNAL SEPARATION AND ACOUSTIC ECHO CANCELING

Daniël W.E. Schobben and Piet C.W. Sommen

Eindhoven University of Technology,
P.O. Box 513, 5600 MB Eindhoven, The Netherlands
Email: D.W.E.Schobben@ele.tue.nl
URL: http://www.esp.ele.tue.nl/˜daniels/

ABSTRACT

The problem of joint blind signal separation and acoustic echo cancelling arises in applications such as teleconferencing and voice controlled machinery. Microphones pick up a signal of the desired speaker together with contributions of other speakers and loudspeakers in these applications. The contributions of these loudspeaker signals to the microphone signals need to be cancelled. The remaining signals are then separated so that the individual local speakers are recovered.

In this paper an extension of the recently introduced Convolutive Blind Signal Separation algorithm; CoBliSS is presented. This extended algorithm is capable of performing combined blind signal separation and acoustical echo cancelling at a low computational cost. The performance of the extended CoBliSS algorithm is evaluated using audio that is recorded in a real acoustical environment.

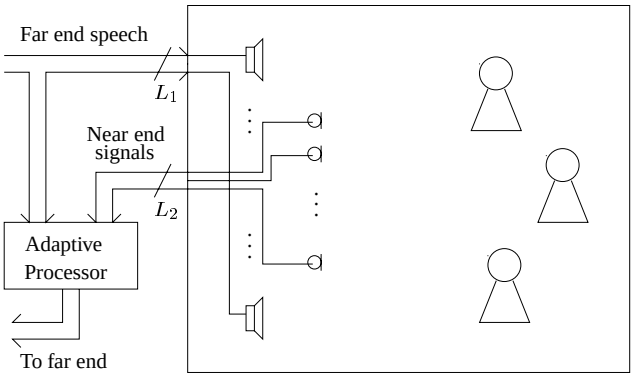

Figure 1: Teleconferencing setup

1. INTRODUCTION

Both Blind Signal Separation (BSS) and Acoustic Echo Canceling (AEC) is required for high quality audio applications such as teleconferencing and voice controlled machinery. The teleconferencing setup is depicted in Figure 1. Typically, both local speakers and reproduced far end sounds or music are present. The adaptive processor can consist of separate BSS and AEC. Recently, several convolutive BSS algorithms have been introduced that are based on Second Order Statistics (SOS) [1, 2, 3, 4, 5, 6]. As conventional AEC's employ SOS too, it seems feasible to merge such BSS and AEC algorithms. Higher quality and lower computational complexity can be achieved in this way [7].

Recently, the CoBliSS BSS algorithm was introduced which is entirely based on SOS [6]. In this paper an extension of CoBliSS is presented which is capable of performing joint AEC and BSS.

The problem of recovering independent signals from mixtures of them which are contaminated with acoustic echos is depicted schematically in Figure 2. The sources $s_1 \dots s_{L_2}$ are the unknown sources, e.g. the local speakers. The sources $s_{L_2+1} \dots s_J$ are the known sources, e.g. far end speech which is reproduced in the same room using loudspeakers. The multi-channel room impulse response is modeled by H. The microphone signals are $x_1 \dots x_{L_2}$. The correlation estimator measures the cross-correlations among all microphone signals and the known sources. This information is used to update the multi-channel filter w which produces $y_1 \dots y_{L_2}$ as outputs. These outputs are the estimates of the unknown sources $s_1 \dots s_{L_2}$. There are two important advantages of combining AEC and BSS into one algorithm that is based on blind signal separation. Conventional AEC is hampered by the presence of active local speakers, which is known as double talk. BSS however profits from simultaneously active speakers. A second advantage is that the performance of the BSS no longer depends on residual echo signals of the AEC.

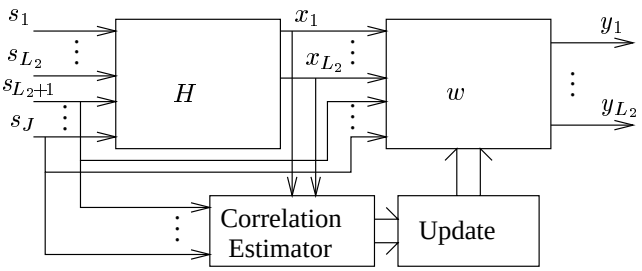

Figure 2: Mixing, unmixing and control system

2. NOTATION

Throughout, time and frequency signals will be denoted by lower case and upper case characters respectively. A character which denotes a vector will be underlined. Superscripts denote the vector or matrix dimensions, a matrix with one superscript is square. Also, A^*, A^T, A^H and A^{-1} denote complex conjugate, matrix transpose, hermitian transpose and matrix inverse respectively and $\jmath^2 = -1$. Element-wise multiplication is denoted by $\otimes$. The expectation operator will be denoted by $E\{.\}$. The $N \times N$ identity matrix and the $K \times L$ zero matrix will be denoted by $\mathbf{I}^N$ and $\mathbf{0}^{K,L}$ respectively. The k,l^{th} element of matrix A and the l^{th} element of vector $\underline{B}$ is denoted as $(A)_{kl}$ and $(\underline{B})_l$ respectively. The $M \times M$ Fourier matrix $\mathcal{F}^M$ is defined as $(\mathcal{F}^M)_{kl} = e^{\frac{-2\jmath\pi\,(k-1)\,(l-1)}{M}}$. The matrix square root sqrtm(.) is defined as $A = \text{sqrtm}(B) \Leftrightarrow A^H A = B$, such that $A^H = A$, with B a complex symmetric matrix, i.e. $B^H = B$. The $N \times N$ mirror matrix $\mathbf{J}^N$ has ones on its anti-diagonal and zeros elsewhere. Time indices are not mentioned explicitly in all equations.

3. DERIVATION

For notational convenience the far end signals are assigned to $x_j = s_j$ for $j = L_2+1 \ldots J$ so that the x_j now represent the input signals for ECoBliSS for $j = 1 \ldots J$. The filter part can now be written for $m = 1 \ldots L_2$

$$y_m[n] = \sum_{l=1}^{J} (\underline{w}_{ml}^N[n])^T \underline{x}_l^N[n]$$

with $\underline{x}_l^N[n] = (x_l[n-N+1] \ldots x_l[n])^T$. The $\underline{w}_{ml}^N[n]$ perform blind signal separation for $l = 1 \ldots L_2$ and acoustical echo canceling for $l = L_2+1 \ldots J$. The idea that forms the basis of ECoBliSS is to make all output signals mutually uncorrelated and uncorrelated with the far end signals.

Similar to the CoBliSS algorithm, blocks of microphone signals are transformed to the frequency domain

$$\underline{X}_a^M = \mathcal{F}^M \begin{pmatrix} x_a[nB-M+1] \\ \vdots \\ x_a[nB] \end{pmatrix}.$$

The blocks are of length M and are overlapping; only B new samples are used per block. The filters are also transformed to the frequency domain

$$\underline{W}_{jc}^M = \mathcal{F}^M \begin{pmatrix} \mathbf{J}^N \underline{w}_{jc}^N \\ \underline{\mathbf{0}}^{M-N} \end{pmatrix}.$$

The transformed microphone signals $\underline{X}_a^M$ and filters $\underline{W}_{jc}^M$ are used to perform the filter operation efficiently in the frequency domain using the overlap save technique [8]. Also, the cross-power estimates are updated efficiently in the frequency domain $\forall a, c$:

$$\underline{P}_{ac}^M := \alpha \underline{P}_{ac}^M + (1-\alpha)\left((\underline{X}_a^M)^* \otimes \underline{X}_c^M\right).$$

The cross-power $\underline{P}_{ac}^M$ is the Fourier transform of the cross-correlation of $x_a[n]$ and $x_c[n]$. The estimation of this cross-power is computationally efficient and fast compared to the estimation of the cross-correlation matrix. The algorithm is based on these cross-powers as there is a known linear relation w between the cross-correlations among the outputs and the cross-correlations among the inputs [4]. This has the advantage that the cross-correlations do not need to be recalculated when the multi-channel unmixing and echo canceling filter w changes.

For every frequency, the filter coefficients and the correlation estimates are grouped in the following way. The p^{th} elements of the vectors $\underline{W}_{ij}^M$ and $\underline{P}_{ij}^M$ are put in a matrix for $i = 1 \ldots L_2, j = 1 \ldots J$ and $i = 1 \ldots J, j = 1 \ldots J$ respectively

$$W_p^{L_2,J} = \begin{pmatrix} (\underline{W}_{11}^M)_p & \ldots & (\underline{W}_{1J}^M)_p \\ \vdots & \ddots & \vdots \\ (\underline{W}_{L_2 1}^M)_p & \ldots & (\underline{W}_{L_2 J}^M)_p \end{pmatrix} = \left(\dot{W}_p^{L_2}\, \ddot{W}_p^{L_2,L_1}\right)$$

$$P_p^J = \begin{pmatrix} (\underline{P}_{11}^M)_p & \ldots & (\underline{P}_{1J}^M)_p \\ \vdots & \ddots & \vdots \\ (\underline{P}_{J1}^M)_p & \ldots & (\underline{P}_{JJ}^M)_p \end{pmatrix}.$$

The $\dot{W}_p^{L_2}$ correspond to the filters that perform the BSS and the $\ddot{W}_p^{L_2,L_1}$ correspond to the filters that perform the AEC. In the CoBliSS algorithm, the weight matrix W_p^J is square and the decorrelation criterion is $(W_p^J)^* P_p^J (W_p^J)^T = \Lambda_p^J$, with Λ_p^J a diagonal matrix. The diagonal elements of Λ_p^J set the power of the recovered signals for frequency p. The off-diagonal zero elements correspond to mutually uncorrelated recovered signals for frequency p. The decorrelation

criterion for ECoBliSS is a generalization of that of CoBliSS

$$\begin{pmatrix} (\dot{W}_p^{L_2})^* & (\ddot{W}_p^{L_2,L_1})^* \\ \mathbf{0}^{L_1,L_2} & \mathbf{I}^{L_1} \end{pmatrix} P_p^J \begin{pmatrix} (\dot{W}_p^{L_2})^T & \mathbf{0}^{L_2,L_1} \\ (\ddot{W}_p^{L_2,L_1})^T & \mathbf{I}^{L_1} \end{pmatrix} = \begin{pmatrix} \Lambda_p^{L_2} & 0^{L_2,L_1} \\ 0^{L_1,L_2} & \breve{P}_p^{L_1} \end{pmatrix}. \quad (1)$$

with $\Lambda_p^{L_2}$ a diagonal constraint matrix as before and $\breve{P}_p^{L_1}$ the cross-power matrix of the known sources for frequency bin p

$$\breve{P}_p^{L_1} = \begin{pmatrix} (P_{L_2+1,L_2+1})_p & \dots & (P_{J,L_2+1})_p \\ \vdots & \ddots & \vdots \\ (P_{J,L_2+1})_p & \dots & (P_{J,J})_p \end{pmatrix}.$$

The weight matrix in (1) consists of the matrices $\dot{W}_p^{L_2}$ and $\ddot{W}_p^{L_2,L_1}$, a zero matrix and a identity matrix. This can be seen as a filter structure that yields both $y_1 \dots y_{L_2}$ and $s_{L_2+1} \dots s_J$ as outputs. The right hand side of (1) prescribes that $y_1 \dots y_{L_2}$ are uncorrelated as Λ_p^J is a diagonal matrix. The zero matrices prescribe that the $y_1 \dots y_{L_2}$ are not correlated with the far end signals $s_{L_2+1} \dots s_J$. The $\breve{P}_p^{L_1}$ is required as the cross-correlations among the far end signals remain unchanged. Equation (1) is rearranged as

$$\begin{pmatrix} (\dot{W}_p^{L_2})^T & \mathbf{0}^{L_2,L_1} \\ (\ddot{W}_p^{L_2,L_1})^T & \mathbf{I}^{L_1} \end{pmatrix} \begin{pmatrix} (\Lambda_p^{L_2})^{-1} & 0^{L_2,L_1} \\ 0^{L_1,L_2} & (\breve{P}_p^{L_1})^{-1} \end{pmatrix} \begin{pmatrix} (\dot{W}_p^{L_2})^* & (\ddot{W}_p^{L_2,L_1})^* \\ \mathbf{0}^{L_1,L_2} & \mathbf{I}^{L_1} \end{pmatrix} = (P_p^J)^{-1} \quad (2)$$

so that

$$\begin{pmatrix} (\dot{W}_p^{L_2})^T(\Lambda_p^{L_2})^{-1}(\dot{W}_p^{L_2})^* & (\dot{W}_p^{L_2})^T(\Lambda_p^{L_2})^{-1}(\ddot{W}_p^{L_2,L_1})^* \\ (\ddot{W}_p^{L_2,L_1})^T(\Lambda_p^{L_2})^{-1}(\dot{W}_p^{L_2})^* & (\ddot{W}_p^{L_2,L_1})^T(\Lambda_p^{L_2})^{-1}(\ddot{W}_p^{L_2,L_1})^* \end{pmatrix} = (P_p^J)^{-1} - \begin{pmatrix} \mathbf{0}^{L_2} & \mathbf{0}^{L_2,L_1} \\ \mathbf{0}^{L_1,L_2} & (\breve{P}_p^{L_1})^{-1} \end{pmatrix}. \quad (3)$$

Note that both sides of this equation are of rank L_2 (with P_p^J and $\breve{P}_p^{L_1}$ full rank)[1] so the weight matrix products are uniquely defined. The inverse of the correlation matrix is now written as

$$(P_p^J)^{-1} = \begin{pmatrix} \dot{P}_p^{L_2} & \ddot{P}_p^{L_2,L_1} \\ \dot{\breve{P}}_p^{L_1,L_2} & \ddot{\breve{P}}_p^{L_1} \end{pmatrix}$$

so that the two relevant equations from (3) are

$$(\dot{W}_p^{L_2})^T(\Lambda_p^{L_2})^{-1}(\dot{W}_p^{L_2})^* = \dot{P}_p^{L_2} \quad (4)$$

$$(\dot{W}_p^{L_2})^T(\Lambda_p^{L_2})^{-1}(\ddot{W}_p^{L_2,L_1})^* = \ddot{P}_p^{L_2,L_1} \quad (5)$$

[1]For the left hand side of (3) this is obvious because it equals the left hand side of (2) with $(\breve{P}_p^{L_1})^{-1}$ replaced by zeros. For the right hand side it can be seen from

$$\text{rank}\left\{\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} - \begin{pmatrix} 0 & 0 \\ 0 & D^{-1} \end{pmatrix}\right\} = \text{rank}\left\{\mathbf{I} - \begin{pmatrix} A & B \\ C & D \end{pmatrix}\begin{pmatrix} 0 & 0 \\ 0 & D^{-1} \end{pmatrix}\right\} = \text{rank}\left\{\begin{pmatrix} \mathbf{I} & -BD^{-1} \\ 0 & 0 \end{pmatrix}\right\}.$$

Similar to CoBliSS the diagonal matrix $(\Lambda_p^{L_2})^{-1}$ can be absorbed by the weight matrices. Ideally, $\Lambda_p^{L_2}$ contains the powers of the sources to be recovered on its diagonal for all frequencies p. As these powers are unknown in practice, $\Lambda_p^{L_2}$ is set equal to the identity matrix. The impact of this is that the spectra of the recovered sources will be flattened. When the signals have almost no energy for a certain frequency, the filter coefficients will be made large so that the recovered output signals have unity power as prescribed by $\Lambda_p^{L_2} = \mathbf{I}^{L2}$. This can be compensated for by normalizing the weight matrices so that they all have the same norm. Therefore, the BSS matrix $\dot{W}_p^{L_2}$ can be initialized by using the matrix square root to decompose the inverted cross-correlation matrix

$$\dot{W}_p^{L_2} = (\text{sqrtm}\{\dot{P}_p^{L_2}\})^*.$$

The AEC matrix $\ddot{W}_p^{L_2,L_1}$ depends on the BSS matrix $\dot{W}_p^{L_2}$. It is found from rewriting (5)

$$\ddot{W}_p^{L_2,L_1} = (\dot{W}_p^{L_2})^{-H}(\ddot{P}_p^{L_2,L_1})^* \quad (6)$$

In this approach the AEC weight matrix is calculated from the BSS weight matrix. This is an advantage over traditional systems with separate BSS and AEC, where the BSS is updated independent of the AEC. When the AEC changes in such a system, the BSS has to re-converge.

When the weights are initialized, the cross-correlation estimates are updated and the weights are constraint so that they correspond to linear FIR filters of length N:

$$(\mathcal{F}^M)^{-1}\underline{W}_{jc}^M = \begin{pmatrix} \underline{w}_{jc}^N \\ \underline{\mathbf{0}}^{M-N} \end{pmatrix}. \quad (7)$$

This is done by replacing the coefficients that should equal zero by zeros in the time domain. Equations (4) and (5) no longer hold after these modifications. The objective is now to modify the weight matrices so that these equations hold again. These modifications must be as small as possible to ensure that the modified weight matrices still approximately satisfy (7). Similar to CoBliSS, this is done by post-multiplying the $\dot{W}_p^{L_2}$ by $C_p^{L_2}$, with

$$C_p^{L_2} = (\text{sqrtm}\{(\dot{W}_p^{L_2})^T(\dot{W}_p^{L_2})^*\}^*)^{-1}\text{sqrtm}\{(\dot{P}_p^{L_2})^{-1}\}^*.$$

The AEC matrix can again be found from (6). This procedure is repeated continuously. Together these steps form the ECoBliSS algorithm which is summarized in the next section.

4. ALGORITHM OVERVIEW

In this section an overview is given of the ECoBliSS algorithm. The ECoBliSS algorithm consists of the following steps;

1. Blocks of input data are transformed to the frequency domain for $a = 1 \ldots J$
$$\underline{X}_a^M[nB] = \mathcal{F}^M \begin{pmatrix} x_a[nB-M+1] \\ \vdots \\ x_a[nB] \end{pmatrix}.$$
These blocks are of length M and are overlapping; only B new samples are used per block.

2. Cross-power estimates are updated efficiently in the frequency domain $\forall a, c$:
$$\begin{aligned} \underline{P}_{ac}^M[nB] &:= \alpha \underline{P}_{ac}^M[(n-1)B] \\ &+ (1-\alpha)\left((\underline{X}_a^M)^*[nB] \otimes \underline{X}_c^M[nB]\right). \end{aligned}$$
The forgetting factor α may vary from 0 to 1 depending on the application. Usually α is chosen near to 1, e.g. $\alpha = 0.99$. In contrast to CoBliSS, these cross-powers are calculated among both the microphone signals and the far end signals.

3. When the cross-correlation matrices are initially estimated the weights that correspond to the signal separation are initialized using the matrix square root $\forall p : \dot{W}_p^{L_2} = \text{sqrtm}(\dot{P}_p^{L_2})^*$. The weights that correspond to the acoustical echo canceling are calculated $\forall p : \ddot{W}_p^{L_2,L_1} = ((\dot{W}_p^{L_2})^H)^{-1}(\ddot{P}_p^{L_2,L_1})^*$.

4. The weights are changed so that (1) holds again $\forall p : \dot{W}_p^{L_2} := \dot{W}_p^{L_2} C_p^{L_2}$
with $C_p^{L_2} = \text{sqrtm}((\dot{W}_p^{L_2})^H \dot{W}_p^{L_2})^{-1} \text{sqrtm}(B_p^J)^*$
Note: This step can be omitted if the initialization has just been done.

5. The weight matrices are normalized using the 2-norm
$\dot{W}_p^{L_2} := \frac{\dot{W}_p^{L_2}}{\left|\left|\begin{pmatrix} \dot{W}_p^{L_2} & \ddot{W}_p^{L_2,L_1} \end{pmatrix}\right|\right|}$
and $\ddot{W}_p^{L_2,L_1} := \frac{\ddot{W}_p^{L_2,L_1}}{\left|\left|\begin{pmatrix} \dot{W}_p^{L_2} & \ddot{W}_p^{L_2,L_1} \end{pmatrix}\right|\right|}$.
Note that $\begin{pmatrix} \dot{W}_p^{L_2} & \ddot{W}_p^{L_2,L_1} \end{pmatrix}$ means that $\dot{W}_p^{L_2}$ and $\ddot{W}_p^{L_2,L_1}$ are placed next to each other in one matrix.

6. The weights are constraint so that the frequency domain weights corresponds to time domain filters of length $N, \forall p$:
$\underline{W}_{jc}^M := \mathcal{F}^M \begin{pmatrix} \mathbf{I}^N & \mathbf{0}^{N,M-N} \\ \mathbf{0}^{M-N,N} & \mathbf{0}^{M-N} \end{pmatrix} (\mathcal{F}^M)^{-1} \underline{W}_{jc}^M$.
Note that the $(W_p^J)_{ac} = (\underline{W}_{ac}^M)_p$.

7. The filtering is performed efficiently in the frequency domain using the overlap-save method [8] to ideally obtain the separated and echo free outputs
$\underline{y}_j^B = (\mathbf{0}^{B,M-B}\mathbf{I}^B)(\mathcal{F}^M)^{-1} \sum_{a=1}^J (\underline{X}_a^M \otimes \underline{W}_{ja}^M)$.

8. All steps are repeated iteratively except for the initialization in item 3.

Note that similar to CoBliSS, the filtering and the weight update can be calculated independently. When the update of the cross-correlations is slower than the weight update for example, reducing the weight update rate lowers the computational complexity at the cost of only a slightly slower convergence of the system.

5. EXPERIMENTS

Experiments were done with audio recorded in a real acoustical environment. The room that is used for the recordings has dimensions 3.4 x 3.8 x 5.2 m (height x width x depth) and is depicted in Figure 3. Two persons read 4 sentences aloud. Far end speech was introduced by playing the French radio news over a small loudspeaker. The resulting sound was recorded by two microphones which were spaced 58 cm apart. The recordings are sampled at 24kHz, with 16 bit accuracy. The separation filters are of length 512 tabs. In an initial experiment, the CoBliSS algorithm is used. The microphone signals and the far end signal used as inputs for the algorithm. For the sake of computational complexity, only one update per every 2560 samples is done. In this experiment, the algorithm converges to a good solution within 0.25 second. This convergence time is required to make good estimates of the cross-correlations.

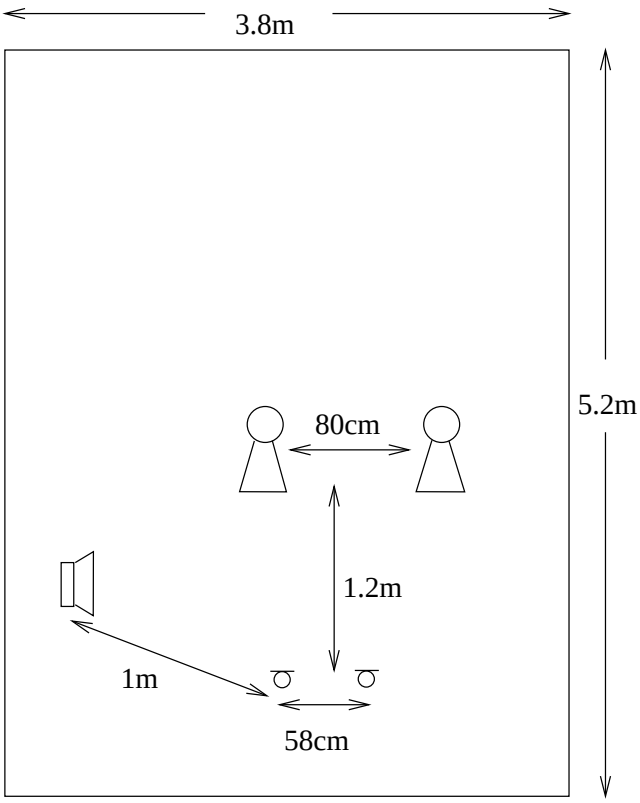

Figure 3: Recording setup

Therefore, increasing the update rate does not significantly increase the convergence time. The resulting sound files can be played online[2]. The three outputs of the CoBliSS algorithm indeed consist of the two separated speakers and the far end speech which is contaminated with some crosstalk. An important advantage of the CoBliSS over a system that uses conventional AEC is that it performs acoustic echo canceling which is not impaired by double talk situations. This makes it suitable for applications like teleconferencing, hands free telephony, etc. A drawback is that blind signal separation is based on little information so that it becomes becomes much more difficult in a real environment when the number of sources increases.

The same experiment is repeated with ECoBliSS. The ECoBliSS algorithm does exploit the fact that the far end signal readily is a source and recovers the unknown speakers only. The outputs of ECoBliSS sound significantly better than those of the first experiment. Also, the computational complexity of extended CoBliSS is significantly lower, as the number of filters decreases and the matrices that need to be decomposed become smaller. The signals that are recovered by ECoBliSS in this experiment can also be played online[2].

6. CONCLUSIONS

An extension of the recently introduced CoBliSS BSS algorithm is presented in this paper. This extension is capable of performing joint acoustic echo canceling and blind signal separation at a low computational complexity. The performance is verified using data that is collected in a real world environment. The extended algorithm exhibits a performance that is superior to that of the CoBliSS algorithm when the known signal is used as an additional input. The computational complexity of the extended CoBliSS algorithm is significantly lower however which is important for real-time operation. An additional advantage of the CoBliSS algorithm over conventional echo canceling is that it can operate in double talk situations without complications. This makes it suitable for applications such as teleconferencing and hands free telephony.

7. REFERENCES

[1] S. Ikeda and N. Murata. A method of ica in time-frequency domain. *Proc. Int. Workshop Independent Component Analysis and Blind Signal Separation*, pages 365–370, Jan. 1999.

[2] L. Parra and C. Spence. Convolutive blind source separation based on multiple decorrelation. In *Proc. of NNSP98*, Cambridge, UK, September 1998.

[3] S. v. Gerven. *Adaptive noise cancellation and signal separation with applications to speech enhancement*. Katholieke Universiteit Leuven, March 1996. Ph.D. Thesis.

[4] D.C.B. Chan. *Blind signal separation*. Cambridge: Thesis University of Cambridge, 1997. Ph. D. Thesis.

[5] E. Weinstein, M. Feder, and A. V. Oppenheim. Multi-channel signal separation by decorrelation. *IEEE Trans. on Speech and Audio Processing*, 1(4):405–413, Oct. 1993.

[6] D.W.E. Schobben and P.C.W. Sommen. A new convolutive blind signal separation algorithm based on second order statistics. In *Proc. Int. Conf. on Signal and Image Processing*, pages 564–569, Oct. 1998.

[7] D.W.E. Schobben and P.C.W. Sommen. Transparent communication. In *Proc. IEEE Benelux Signal Processing Chapter Symposium*, pages 171–174, Mar. 1998.

[8] D.W.E. Schobben, G.P.M. Egelmeers, and P.C.W. Sommen. Efficient realization of the block frequency domain adaptive filter. In *Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2257–2260, Apr. 1997.

[2] http://www.esp.ele.tue.nl/~daniels/