Transparent communication

D.W.E. Schobben and *P.C.W. Sommen*

Electrical Engineering Department
Eindhoven University of Technology
P.O. Box 513, 5600 MB Eindhoven, the Netherlands

tel. +31/402472366 fax. +31/402455674
email : D.W.E.Schobben@ele.tue.nl and P.C.W.Sommen@ele.tue.nl

Abstract

Transparent communication refers to the audio signal processing which is applied in communication applications. The goal is to make the audio as transparent as possible in the sense that the reproduced audio should ideally be free from reverberation, noise, acoustical echos and mixed speakers. Application areas are for example teleconferencing and hands-free telephony. This paper presents new ideas for the implementation of such a system. In particular, the use of blind signal separation is examined and new ideas are presented for the joint implementation of the Multi-Channel Acoustical Echo Canceler (MC-AEC) and the Blind Signal Separation (BSS). In this way, acoustical quality can be improved at a reduced computational cost.

1 INTRODUCTION

In a teleconferencing setup, plain recording of near end speech can result in ineligible reproduced speech at the far end. This reproduced speech is observed as a noisy unnatural sounding mixture of multiple speech signals which also contains acoustical echos. Besides this, data compression is far less efficient for such a signal than for clean speech. The degradation of the near end speech recordings is caused by the following:

- Reproduced far end speech propagates towards the microphones and generates acoustical echos.
- Microphones pickup an acoustical mixture of several speech signals
- Microphone signals have reduced signal to noise ratio due to the pickup of unwanted surrounding noise.
- Speech signals are affected by the acoustical reverberation.

Quality can be improved if multiple microphones are used. Digital signal processing is applied to these microphone signals in order to ideally separate the speakers, cancel the acoustical echos, eliminate the reverberation and suppress surrounding noise. Figure 1 depicts a teleconferencing setup, in which L_1 loudspeakers reproduce the far end speech and L_2 microphones pickup degraded near end speech. An

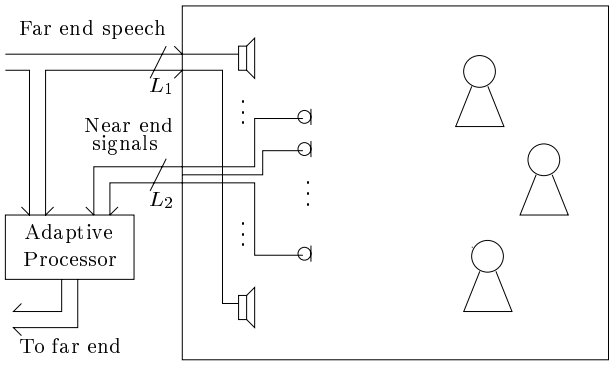

Figure 1: Teleconferencing setup

approach is presented in this paper to combine the BSS and MC-AEC that are needed to achieve transparent communication. In this way, both performance and computational cost can be improved. The problem of dereverberation is not addressed in this paper.

2 BLIND SIGNAL SEPARATION

The goal of blind signal separation is to recover estimates of the sources signal from an observed mixture of them. Figure 2 depicts the mixing and unmixing system in this context. The mixing system H can be modeled by FIR filters that are present between every input and every output of this multichannel system. For acoustical applications these filters can have a length of several thousands of tabs, depending on the sample rate and properties of the room in which the microphones are placed. The goal of the unmixing system is to produce outputs that are linear functions of the sources, $y_i = f_i(s_i), 1 \leq i \leq J$, with J the number of inputs and outputs of the mix-

ing and unmixing system. Note that the permutation of the recovered signals and the linear functions f_i are ambiguous when no properties of the sources themselves or their locations are used. In practical situations these permutations are often not important. Also, the unmixing system can be restricted so that it has amplitude responses that are relatively flat. In this way, its outputs sound just as natural as its inputs.

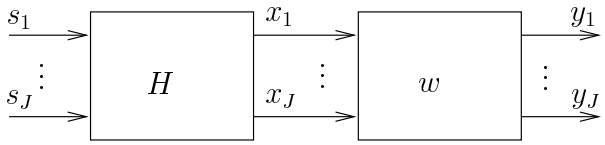

Figure 2: Blind signal separation

From the separated speech signals, the strongest one may be sent to the far end. Audio can be made even more transparent by sending more than one speech signal and play them via several distinct loudspeakers. If only one speech signal is required, it is also possible to track only the strongest one. This can be done using array signal processing [1, 2, 3]. Array signal processing typically assumes knowledge of the geometry of the array and tracks the strongest source. Tracking only the strongest sources has the disadvantage that it gives poor performance at the time that it switches from one source to the other. Blind signal separation is usually based on the fact that the speech signals are independent of each other and typically tries to recover all source signals. Even if only one speech signal is sent to the far end, recovering all sources has the advantage that the strongest one can be picked from the outputs at all times, without the system having to reconverge. The unmixing system w consists of a set of J^2 FIR filters similar to the mixing system. For blind signal separation, these filters are controlled to minimize a cost function [4, 5, 6]. This cost function can be based, for example, on mutual information, maximum likelihood, second or higher order statistics. A priory knowledge of the probability density functions (pdf's) of the speech signals can be used as a tool to adaptively maximize the mutual information among the outputs of the BSS scheme [7, 8, 9].
Another interesting approach which will be used in this paper is to minimize cross-correlations among the outputs of the BSS scheme [10]. This approach does not require any a priory knowledge other than the statistical independence of the speech signals. The objective of this approach can be given by

$$\min_{w} \sum_{l} \sum_{i} \sum_{j \neq i} |r_{y_i y_j}[l]| \qquad (1)$$

This cost function will be small when the filtercoefficients w are chosen such that the outputs of the BSS become independent of each other in terms of their second order statistics. The correlation lags l which are used in this cost function must be from $-N+1$ to $N-1$, with N the length of the FIR filters in the unmixing system. This ensures that the problem is not ambiguous. Using more lags is also allowed and can further improve performance [10]. The crosscorrelation $r_{y_i y_j}[l]$ can be expressed in the crosscorrelation of the input of the BSS $r_{x_i x_j}[l]$

$$\begin{aligned} r_{y_i y_j}[l] &= E\{y_i[n] y_j[n+l]\} \\ &= \sum_{a=1}^{L_2} \sum_{c=1}^{L_2} \sum_{b=0}^{N-1} \sum_{d=0}^{N-1} w_{ia}[b] w_{jc}[d] r_{x_a x_c}[l+b-d] \end{aligned} \qquad (2)$$

In this notation, $w_{ia}[b]$ is the b^{th} tab of the FIR-filter which is present between the a^{th} input and the i^{th} output of the BSS. The advantage of (2) over (1) is that it no longer explicitly contains $r_{y_i y_j}[l]$ which changes when the filter coefficients change. Instead, $r_{x_i x_j}[l]$ can be estimated once from a data set, and the filter coefficients can be found from minimizing the cost function which is expressed in $r_{x_i x_j}[l]$ and in the filter coefficients only.

3 COMBINING MC-AEC & BSS

First, two traditional approaches will be presented in this section. It will be argued that they have important drawbacks which cannot be solved by applying AEC and BSS independently.

3.1 Separate BSS & MC-AEC

The acoustical echos (i.e. the far end speech signals) that are picked up by the microphone array can be cancelled using a MC-AEC [11]. Next, BSS is applied to separate the near end speech signals. This is depicted in Figure 3. This approach has the following drawbacks

- The MC-AEC is not able to work well with double talk, i.e. when there is both near end speech and far end speech at the same time. This is a problem when tracking time varying acoustical transfer functions. The overall performance of the system will degrade since the performance of the BSS depends on the performance of the MC-AEC.

- The separate implementation of MC-AEC and BSS can result in a considerable computational workload. This is especially true for cases with several loudspeakers and many microphones but where only a few outputs need to be retrieved.

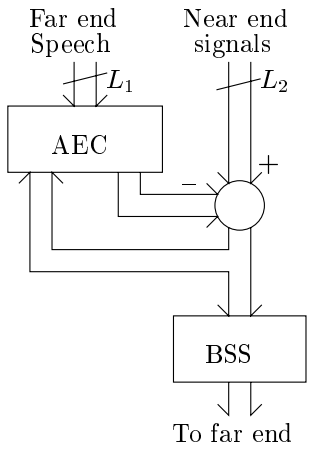

Figure 3: Adaptive echo canceling followed by blind signal separation

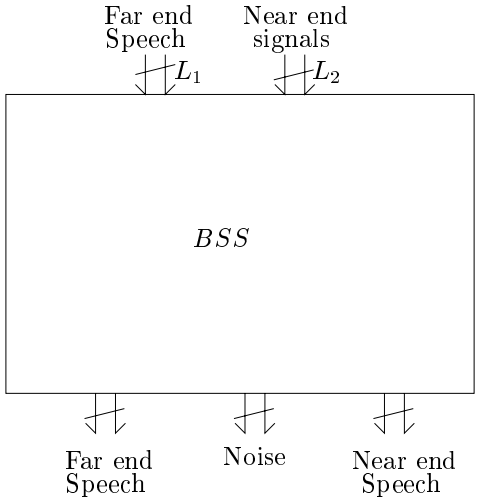

Figure 4: Combined adaptive echo canceling and blind signal separation

3.2 BSS without MC-AEC

Theoretically, the BSS can classify acoustical echos as sources that are independent of the near end speech signals. Therefore, a possible solution could be to use $L_1 + L_2$ microphones and let the BSS retrieve the far end speech from this as outputs which are independent of the retrieved near end speech. Simulations showed however that the BSS is not able to do this with an accuracy comparable to that of the MC-AEC. Furthermore, the separation becomes more difficult when more sources are involved. In the following section, an approach is presented which also makes use of the far end speech itself. The performance of the system is greatly improved by this.

4 JOINT BSS & MC-AEC

In order to obtain a succesfull combination of BSS and MC-AEC, the objective function (1) is extended to

$$\min_{w} \sum_{l} \sum_{i} \left(\sum_{j \neq i} |r_{y_i y_j}[l]| + \sum_{m} |r_{z_m y_i}[l]| + |r_{y_i z_m}[l]| \right), \quad (3)$$

with z_m the m^{th} far end speech signal. In this way, the BSS will produce outputs which are not only independent of each other, but they are also independent of the far end speech signals. In fact, the BSS is extended by adding the far end speech as input signals. The BSS with its inputs and outputs is depicted in Fig. 4. So, when the system is controlled by optimizing (3), it can be considered as a special case of optimizing (2) with the restriction that far end speech must also appear as outputs of the BSS. Therefore the filters between the far end speech input and outputs are fixed to the unit impulse response. Furthermore, the filters from the microphone inputs to the far end speech outputs are kept identically equal to zero.

Besides the trivial far end speech outputs, the BSS produces separated near end speech which are independent of the far end speech. In this way acoustical echos are suppressed. The number of microphones used in this approach must be greater than or equal to the number of local speakers. If the number of microphones is larger than the number of local speakers, the BSS will also generate noise outputs which correspond to noisy observations of the local speakers, or strong physical noise sources such as the fan of an overhead projector.

5 EXPERIMENTS

Experiments were carried out with audio signals recorded in a real acoustical environment. The room which is used for the recordings was 3.4 x 3.8 x 5.2 m (height x width x depth). Two live speakers read 4 sentences aloud. Also, far end speech was introduced by playing prerecorded French news over a small loudspeaker. The resulting sound was recorded by two microphones. The setup is depicted in Figure 5. The microphone signals and the far end speech were used to minimize the extended objective function (3) off-line. The FIR filters that were used in the BSS all have 512 tabs. All signals were sampled at 24KHz with a 16 bit accuracy. The output of the BSS shows clear separation of the speech and good suppression of the acoustical echos. An AEC could however be used to further remove the residual echos. For this application area, quality can not be expressed in terms of SNR in a straightforward way because the separated speech signals don't resemble

the original speech signals, but are linear functions of them instead. Both the original speech signals and the linear functions are unknown. In order to give an impression of the improvements that can be achieved using this approach, the microphone signals, far end speech, and the output of the BSS are available for listening. The tracks can be found in WAV-format at

`http://www.ses.ele.tue.nl/persons/daniels/`

by choosing "Transparent Communication" from the publication list. There is also BSS page on which the latest research results will be presented.

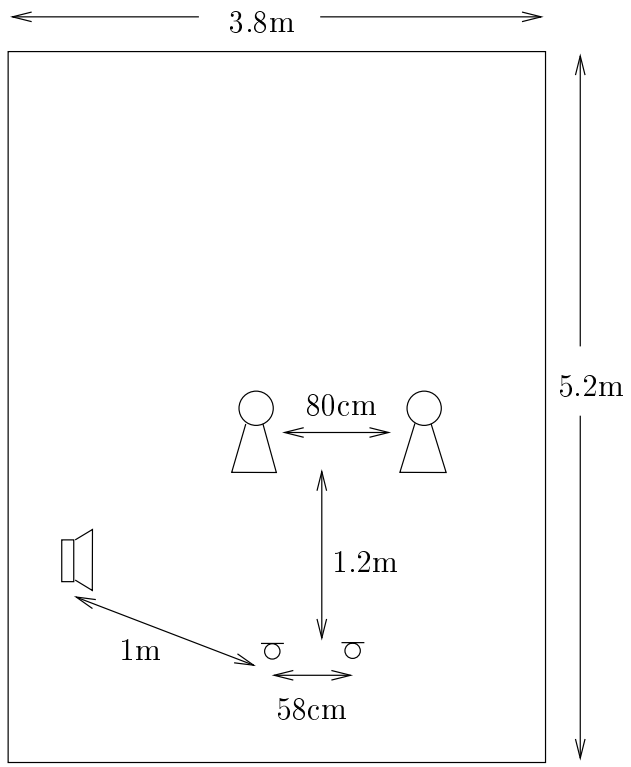

Figure 5: Recording setup

6 Conclusions and future work

Blind signal separation is an important tool in applications like teleconferencing. In this paper, the concept of blind signal separation is extended to incorporate acoustical echo cancellation. Experiments with real acoustical measurements show that this extended approach exhibits a good performance. Subject to further study is the online (adaptive) implementation of the extended algorithm. Important issues to be considered are the computational workload and the convergence speed.

References

[1] K.M. Buckley B.D. van Veen. Beamforming: A versatile approach to digital filtering. *AASP Mag.*, 5(2):4–24, 1988.

[2] Y. Grenier S. Affes. A speaker tracking array of microphones. *IEEE Trans. on Speech and Audio Proc.*, 5:425–437, Sept. 1997.

[3] W. Kellerman. Strategies for combining acoustic echo cancellation and adaptive beamforming microphone arrays. In *Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, pages 219–222, Apr. 1997.

[4] P. Smaragdis. Blind separation of convolved mixtures in the frequency domain. In *Int. Workshop on Independence & Artificial Neural Networks*, Febr. 1998.

[5] R. Lambert T-W. Lee, A.J. Bell. Blind separation of delayed and convolved sources. *Advances in Neural Inf. Proc. Systems 9*, pages 758–764, 1997.

[6] A.J. Bell R.H. Lambert. Blind separation of multiple speakers in a multipath environment. In *Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, Apr. 1997.

[7] T.J. Sejnowski A.J. Bell. An information maximisation approach to blind separation and blind deconvolution. *Neural Computation 7*, pages 1129–1159, 1995. MIT Press, Cambridge MA.

[8] R.H. Lambert. *Multichannel Blind Deconvolution: FIR Matrix Algebra and Separation of Multipath Mixtures.* University of Southern California, May 1996. Ph.D. Thesis.

[9] K. Torkkola. Blind separation of delayed sources based on information maximization. In *IEEE Workshop on Neural Networks for Signal Proc.*, Sept. 1996.

[10] D.C.B. Chan. *Blind Signal Separation.* Cambridge: Thesis University of Cambridge, 1997. Ph. D. Thesis.

[11] Y. Grenier F. Alberge, P. Duhamel. A combinded fdaf/wsaf algorithm for stereophonic acoustic echo cancellation. In *Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, May 1998.