AN ACOUSTIC HUMAN-MACHINE INTERFACE WITH MULTI-CHANNEL SOUND REPRODUCTION

H. Buchner and W. Kellermann*
Telecommunications Laboratory
University of Erlangen-Nuremberg
Cauerstr. 7, D-91058 Erlangen, Germany

Abstract - For hands-free man-machine audio interfaces with multi-channel sound reproduction and automatic speech recognition (ASR), sometimes both an acoustic echo canceller (AEC) and a beamforming (BF) microphone array are necessary for sufficient recognition rates. In the context of multimedia systems, multi-channel sound reproduction (e.g. stereo or 5.1 channel - surround systems) typically requires multi-channel AEC (M-C AEC). With M-C AEC being known to be a very demanding problem in signal processing, no practically relevant simulation results have yet been presented for more than two channels. In this contribution, we examine a recently proposed frequency-domain adaptive filtering scheme [5] with some extensions for the case of more than two loudspeaker channels. The results using this approach show a remarkable performance at a relatively moderate computational complexity. Moreover, we show how to efficiently combine this M-C AEC structure with microphone arrays.

INTRODUCTION

The current progress in signal processing hardware, electronic communications technologies and increased functionality of multimedia systems goes hand in hand with a demand for more intuitive hands-free interfaces between the user and the system. However for speech control applications, typical acoustic environments (*Fig.* 1) are still very demanding for state-of-the-art ASR because of three factors: (i) interfering noise or other speakers, (ii) the reverberation of the speech to be recognized and (iii) the high-volume sound produced by the system itself.

An effective approach to (i) and (ii) is the use of a microphone array directing a beam of increased sensitivity at the active talker. The steerable beam may be controlled by some (multimodal) speaker localization. In order to achieve a sufficient word accuracy during sound production of the system, an echo canceller is necessary to cope with problem (iii). AEC is the only

*This work was supported by a grant from Grundig AG, Nuremberg leading the EMBASSI consortium for seamless man-machine interfaces (`http://www.embassi.de`).

known method which has the potential to remove even high volume feedback of loudspeaker signals to a microphone (*Fig.* 2) without impairing the desired speech signals.

Here, in the context of advanced multimedia systems, multi-channel sound reproduction (e.g. stereo or 5.1 channel - surround systems) requires multi-channel AEC.

The problems of M-C AEC include all those known for mono AEC (e.g. [1]), but in addition to that, M-C AEC has to cope with high correlation of the different loudspeaker signals, which in turn cause correlated echoes which cannot easily be distinguished in the microphone signal [2]. The correlation results from the fact that the signals are almost always derived from common sound sources in the transmission room (e.g. a TV studio, *Fig.* 2). Due to the very ill-conditioned problem, more complex algorithms taking the cross-correlation into account have to be applied. Until recently, even the step from mono AEC to the stereo case with only one microphone signal to be compensated has been considered to be very difficult in terms of performance and computational complexity. In the following section we investigate and extend a recently proposed and promising frequency-domain scheme [5] for an increased number of loudspeaker channels.

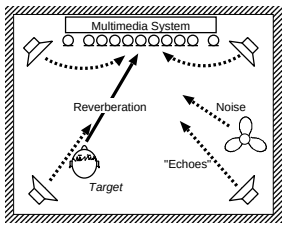

Figure 1: Interface with sound reproduction

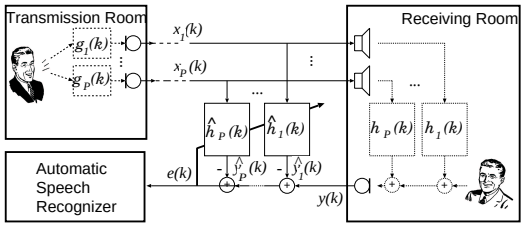

Figure 2: Conventional M-C AEC structure and ASR

M-C AEC IN THE FREQUENCY DOMAIN

Frequency-domain adaptive filtering (FDAF) is well known for its very low complexity and increased convergence speed when properly designed. However, only recently a rigorous derivation of FDAF allowing a straightforward and powerful generalization to the multi-channel case has been presented [5]. In the following considerations we assume a block size equal to the filter length. Generalization to a partitioned version is straightforward, but for the speech recognizer a slightly delayed input signal poses no problem.

The two advantages of this structure, low complexity and fast convergence, are based on linear filtering via fast convolution as well as the approximate diagonalization of Toeplitz matrices by the Discrete Fourier Transform (DFT). For the fast convolution we apply here the overlap-save (OLS) method [1].

In order to compensate the P electro-acoustic echo paths (*Fig.* 2) to one microphone (signal $y(k)$) in the frequency domain, we first transform blocks

(length $2L$) of all loudspeaker signals x_i, $i = 1, \ldots, P$ to the discrete Fourier domain,

$$\boldsymbol{X}_i(m) = diag\{\boldsymbol{F}[x_i(m\frac{L}{\alpha} - L + 1), \cdots, x_i(m\frac{L}{\alpha} + L)]^T\}. \quad (1)$$

$\boldsymbol{F}$ denotes the DFT matrix, L is the modeling filter length, m is the block index, and α denotes a factor capturing the overlap of successive blocks which balances the number of iterations (faster convergence) and computational complexity. The echo replicas $\hat{\boldsymbol{Y}}_i(m)$ in the DFT domain are then generated by multiplying the weights $\hat{\boldsymbol{H}}_i(m)$ with the elements of the diagonal matrix $\boldsymbol{X}_i(m)$. The block of residual errors (including an OLS constraint with the data window $\boldsymbol{W} = diag\{[\boldsymbol{0}_{1\times L}\boldsymbol{1}_{1\times L}]\}$ [5]) becomes

$$\tilde{\boldsymbol{e}}(m) = [\boldsymbol{0}_{1\times L} \quad y(m\frac{L}{\alpha}+1), \cdots, y(m\frac{L}{\alpha}+L)]^T - \boldsymbol{W}\boldsymbol{F}^{-1}\sum_{i=1}^{P}\boldsymbol{X}_i(m)\hat{\boldsymbol{H}}_i(m). \quad (2)$$

This vector serves both as the output signal (e.g., to the ASR, last $\frac{L}{\alpha}$ elements of $\tilde{\boldsymbol{e}}(m)$) and the feedback for the next adaptation step in order to identify and track the impulse responses for the models. Using the transformed error vectors $\tilde{\boldsymbol{E}}_b(m) = \boldsymbol{F}\tilde{\boldsymbol{e}}_b(m)$, the filter weights can be adapted using a recursive least-squares error criterion [5]. This leads to the following update equation for $\hat{\boldsymbol{H}} = [\hat{\boldsymbol{H}}_1^T \hat{\boldsymbol{H}}_2^T \cdots \hat{\boldsymbol{H}}_P^T]^T$:

$$\hat{\boldsymbol{H}}(m+1) = \hat{\boldsymbol{H}}(m) + \mu\boldsymbol{S}^{-1}(m)\boldsymbol{X}^H(m)\tilde{\boldsymbol{E}}(m). \quad (3)$$

This equation is a good approximation to the well-known recursive least-squares (RLS) algorithm in the time domain (e.g. [1, 4]) which provides optimum convergence speed at very high cost. The product

$$\boldsymbol{K}(m) := \boldsymbol{S}^{-1}(m)\boldsymbol{X}^H(m) \quad (4)$$

in Eq. (3) is the frequency-domain analogon to the *Kalman gain vector* as known from the RLS algorithm [4]. Here, however the matrices $\boldsymbol{S}$ to be inverted (cross power spectra) and $\boldsymbol{X} = [\boldsymbol{X}_1\boldsymbol{X}_2\cdots\boldsymbol{X}_P]$ are block diagonal in the *unconstrained* version [5] of the algorithm, i.e. Eq. (4) is the solution of a $P \times P$ system of linear equations of block matrices. This makes it possible to decompose Eq. (3) into P single-channel update equations

$$\hat{\boldsymbol{H}}_i(m+1) = \hat{\boldsymbol{H}}_i(m) + \mu\boldsymbol{K}_i\tilde{\boldsymbol{E}}(m). \quad (5)$$

with modified Kalman gains $\boldsymbol{K}_i(m)$ taking the cross-correlations into account. These decomposed update equations can then be calculated element-wise and the (cross) power spectra are estimated by the simple recursive formulae

$$\boldsymbol{S}_{i,j}(m) = \lambda\boldsymbol{S}_{i,j}(m-1) + (1-\lambda)\boldsymbol{X}_i^*(m)\boldsymbol{X}_j(m), \quad (6)$$

where $i, j = 1, \ldots, P$ and $\boldsymbol{S}_{j,i}(\cdot) = \boldsymbol{S}_{i,j}^*(\cdot)$. λ is a forgetting factor [1]. For the two-channel case, the update equations can be easily written in an explicit form [1], e.g. for the first channel, we have

$$\begin{aligned}\hat{\boldsymbol{H}}_1(m+1) &= \hat{\boldsymbol{H}}_1(m) + \mu \boldsymbol{S}_1^{-1}[\boldsymbol{X}_1^*(m) - \boldsymbol{S}_{1,2}\tilde{\boldsymbol{S}}_{2,2}^{-1}\boldsymbol{X}_2^*(m)]\tilde{\boldsymbol{E}}(m), \qquad (7)\\ \boldsymbol{S}_i(m) &= \tilde{\boldsymbol{S}}_{i,i}(m)[\boldsymbol{I}_{2L\times 2L} - \boldsymbol{S}_{1,2}^*(m)\boldsymbol{S}_{1,2}(m)\{\tilde{\boldsymbol{S}}_{1,1}(m)\tilde{\boldsymbol{S}}_{2,2}(m)\}^{-1}]\end{aligned}$$

For robust adaptation the power spectral densities $\boldsymbol{S}_{i,i}$ must be regularized for these equations. With increasing number of channels this point becomes crucial. We propose a *bin-selective dynamical regularization* according to

$$\boldsymbol{\delta}(m) = \delta_{max}[\exp(-S_{i,i,0}(m)/S_0), \cdots, \exp(-S_{i,i,2L-1}(m)/S_0)]^T. \qquad (8)$$

This method yields improved results compared to fixed regularization or the popular approach of choosing the maximum out of the respective component of $\boldsymbol{S}_{i,i}$ and a fixed threshold δ_{th}.

The solutions of Eq. (4) for more channels can be brought in a form similar to Eq. (7) (e.g. using Cramer's rule). For three channels, we have (omitting, for simplicity, the index m of all matrices)

$$\begin{aligned}\boldsymbol{K}_1 &= \boldsymbol{D}^{-1}[\boldsymbol{X}_1^*(\boldsymbol{S}_{2,2}\boldsymbol{S}_{3,3} - \boldsymbol{S}_{3,2}\boldsymbol{S}_{2,3}) - \boldsymbol{X}_2^*(\boldsymbol{S}_{1,2}\boldsymbol{S}_{3,3} - \boldsymbol{S}_{1,3}\boldsymbol{S}_{3,1}) -\\ &\quad -\boldsymbol{X}_3^*(\boldsymbol{S}_{1,3}\boldsymbol{S}_{2,2} - \boldsymbol{S}_{1,2}\boldsymbol{S}_{2,3})], \qquad (9)\\ \boldsymbol{D} &= \boldsymbol{S}_{1,1}(\boldsymbol{S}_{2,2}\boldsymbol{S}_{3,3} - \boldsymbol{S}_{3,2}\boldsymbol{S}_{2,3}) - \boldsymbol{S}_{2,1}(\boldsymbol{S}_{1,2}\boldsymbol{S}_{3,3} - \boldsymbol{S}_{1,3}\boldsymbol{S}_{3,1}) -\\ &\quad -\boldsymbol{S}_{3,1}(\boldsymbol{S}_{1,3}\boldsymbol{S}_{2,2} - \boldsymbol{S}_{1,2}\boldsymbol{S}_{2,3})\end{aligned}$$

as the first of the three Kalman gain components with the common factor $\boldsymbol{D}$. This representation of Eqs. (7) and (9) can be interpreted as a correction of the interchannel-correlations in $\boldsymbol{K}_i$ between $\boldsymbol{X}_i^*$ and the other input signals. However, for a *practical implementation* of a system with increased number of channels we recommend a more efficient method to calculate Eq. (4). Due to the block diagonal structure of this equation, it can be trivially decomposed in $2L$ equations with $P \times P$ unitary and positive definite matrices $\boldsymbol{S}^{(\nu)}$ for the components $\nu = 1, \ldots, 2L$ on the diagonals. A well-known and stable method for this type of problems is the Cholesky decomposition of $\boldsymbol{S}^{(\nu)}$ followed by solution via backsubstitution, e.g. [6].

For very high number of loudspeakers (e.g. aiming towards actual sound field synthesis for virtual reality) there is an even more efficient option of jointly estimating the *inverse* power spectra $(\boldsymbol{S}^{(\nu)})^{-1}$ and Kalman components $\boldsymbol{K}^{(\nu)}$ in a recursive manner by applying the matrix-inversion lemma (e.g. [4]) to this reduced problem. However, periodic re-initialization of $(\boldsymbol{S}^{(\nu)})^{-1}$ for regularization is necessary to ensure numerical stability. With the intermediate vectors $\boldsymbol{\kappa}^{(\nu)}(m)$ and the corresponding "Riccati equation"

$$\begin{aligned}\boldsymbol{\kappa}^{(\nu)}(m) &= \frac{\lambda^{-1}(\boldsymbol{S}^{(\nu)}(m-1))^{-1}\boldsymbol{X}^{(\nu)H}(m)}{(1-\lambda)^{-1} + \lambda^{-1}\boldsymbol{X}^{(\nu)}(m)(\boldsymbol{S}^{(\nu)}(m-1))^{-1}\boldsymbol{X}^{(\nu)H}(m)}\\ (\boldsymbol{S}^{(\nu)}(m))^{-1} &= \lambda^{-1}[(\boldsymbol{S}^{(\nu)}(m-1))^{-1} - \boldsymbol{\kappa}^{(\nu)}(m)\boldsymbol{X}^{(\nu)}(m)(\boldsymbol{S}^{(\nu)}(m-1))^{-1}],\end{aligned}$$

no explicit matrix inversion is necessary for the calculation of the $2L$ vectors $\boldsymbol{K}^{(\nu)}(m) = (\boldsymbol{S}^{(\nu)}(m))^{-1}\boldsymbol{X}^{(\nu)H}(m)$.

COMBINING M-C AEC WITH A BEAMFORMING MICROPHONE ARRAY

We show now how to efficiently combine the presented frequency-domain M-C AEC with a steerable beamforming microphone array (BF).

In [3] different basic strategies are discussed. A single AEC behind the BF is computationally most attractive, however, for time-variant (adaptive and/or steerable) BF in the echo paths this approach is impracticable. On the other hand, AEC for each of the Q single microphones gives optimum performance but may be computationally very demanding. As a third option, decomposition of the BF into time-variant and time-invariant stages has been proposed [3]. In order to allow beamsteering, a set of $B < Q$ fixed beam signals is generated using the Q microphone signals (*Fig.* 3). The fixed beams cover all potential sources of interest. AEC is applied to each beam, before the beams are summed up with time-varying weights ('voting'). External information on the source position via audio-based, video-based or multimodal object localization can be easily incorporated.

The total complexity of the AEC can be greatly reduced further without detrimental effects on their convergence behaviour by decomposing the M−C AEC algorithm into parts that can be shared among all B beams (or Q microphones) and other parts that have to be computed separately for each beam (*Fig.* 4) or microphone. From Eqs. (4) and (5), we see that the introduction of the Kalman gain $\boldsymbol{K}(m)$ gives another major advantage in this context. The Kalman components only depend on the input data, not on the microphone or beam signals. A significant share of the computational complexity is thus required only once for all beams *('beam-independent')*: it includes FFTs of all loudspeaker signals, power spectrum estimation and regularization. Moreover, the double-talk detection of the AEC is also 'beam-independent' if only one microphone is used for it. Once the Kalman gain is known, the *beam-dependent* part, Eq. (5) corresponds to simple mono FLMS updates and filtering. The relative computational savings due to separation increase with increasing number of beams B and loudspeakers P. For illustration, the complexity of the described algorithm in stereo version (separated) is even for several beams (e.g. $B = 7$ for typical setup) much lower than for the well-known time-domain LMS algorithm [4] with a single microphone.

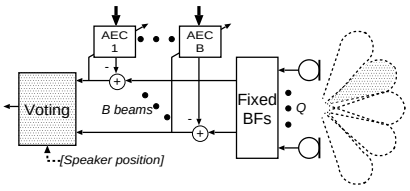

Figure 3: AEC integrated into cascaded beamforming

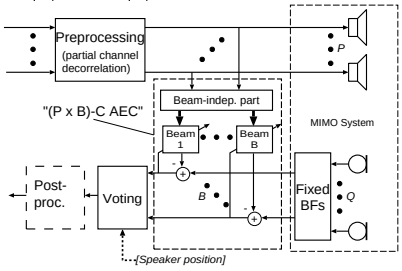

Figure 4: Proposed Structure

SIMULATION RESULTS

For the simulations a speech signal (in the transmission room) was convolved by P different room impulse responses and nonlinearly, but inaudibly preprocessed according to [1] (nonlinearity factor 0.5). The lengths of the receiving room impulse responses were 4096 and the modeling filters were 1024, respectively. White noise for $SNR = 35dB$ was added to the echo on the microphone. *Fig.* 5 shows the misalignment convergence of the described algorithm (solid), $\alpha = 4$ for the multi-channel cases $P = 2, 3, 4, 5$ (from lowest to uppermost line). In *Fig.* 6 the overlap factor α was adjusted to 8 for $P = 3, 4$ and to 16 for $P = 5$. One can clearly see that these lines are then almost indistinguishable. The dashed lines show the corresponding characteristics for the basic NLMS algorithm.

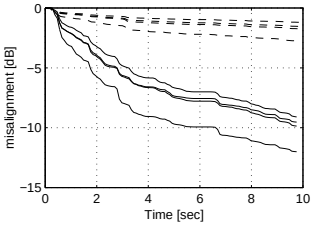

Figure 5: Convergence for P=2,3,4,5 channels, $\alpha = 4$

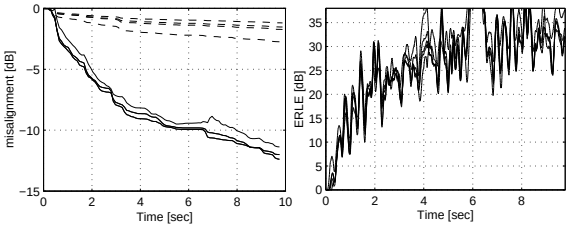

Figure 6: Convergence for $P = 2, 3, 4, 5$ channels and adjusted overlap α

CONCLUSIONS

A fast converging and computationally efficient concept for M-C AEC in conjunction with a steerable beamforming microphone array has been presented. Simulation results have shown that the system also copes very well with more than two loudspeaker channels.

References

[1] S. L. Gay and J. Benesty (eds.), *Acoustic Signal Processing for Telecommunication*, Kluwer Academic Publishers, 2000.

[2] M. M. Sondhi and D. R. Morgan, "Stereophonic Acoustic Echo Cancellation - An Overview of the Fundamental Problem," *IEEE SP Lett.*, vol.2, no.8, pp. 148–151, Aug. 1995.

[3] W. Kellermann, "Strategies for combining acoustic echo cancellation and adaptive beamforming microphone arrays," Proc. ICASSP 1997, pp. 219–222.

[4] S. Haykin, *Adaptive Filter Theory*, 3rd ed., Prentice Hall Inc., Englewood Cliffs, NJ, 1996

[5] J. Benesty and D. R. Morgan, "Frequency-domain adaptive filtering revisited, generalization to the multi-channel case, and application to acoustic echo cancellation," Proc. ICASSP 2000, pp. 789–792.

[6] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 2nd ed., Johns Hopkins, Baltimore, MD, 1989.