AN EFFICIENT COMBINATION OF MULTI-CHANNEL ACOUSTIC ECHO CANCELLATION WITH A BEAMFORMING MICROPHONE ARRAY

H. Buchner, W. Herbordt, and W. Kellermann

University of Erlangen-Nuremberg, Telecommunications Laboratory
Cauerstr. 7, D-91058 Erlangen, Germany
{buchner, herbordt, wk}@LNT.de

ABSTRACT

For hands-free man-machine audio interfaces with multi-channel sound reproduction and automatic speech recognition (ASR), both a multi-channel acoustic echo canceller (M-C AEC) and a beamforming microphone array are necessary for sufficient recognition rates. Based on known strategies for combining single-channel AEC and adaptive beamforming microphone arrays, we discuss special aspects for the extension to multi-channel AEC and propose an efficient system that can be implemented on a regular PC.

1. INTRODUCTION

Acoustic echo cancellation (AEC) is one of the key technologies for hands-free full duplex communication systems where feedback of loudspeaker signals to a microphone occurs (see receiving room in *Fig.* 1). Classical applications are telephony or teleconference systems (suggested by pos. (A) of the switch in *Fig.* 1). With slight modifications this technology can also be adopted as a signal preprocessing unit for automatic speech recognizers in multimedia systems with sound reproduction (pos. (B) in *Fig.* 1). Even during high volume sound output, sufficient recognition rates can be expected with this system. The fundamental idea of any P-channel

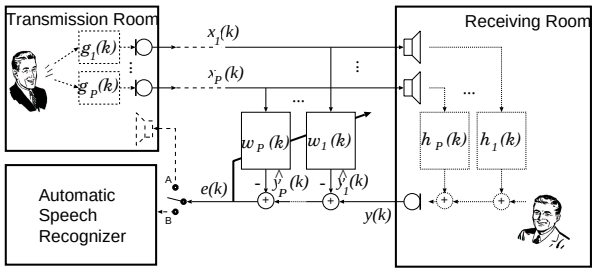

Fig. 1. Conventional M-C AEC structure

AEC structure is to use adaptive FIR filters with impulse response vectors

$$\boldsymbol{w}_i(k) = [w_{i,0}(k), \cdots, w_{i,L-1}(k)]^T, i = 1, \cdots, P, \quad (1)$$

that identify the truncated (generally time-varying) echo path impulse responses

$$\boldsymbol{h}_i(k) = [h_{i,0}(k), \cdots, h_{i,L-1}(k)]^T \quad (2)$$

(*Fig.* 1). The filters $\boldsymbol{w}_i(k)$ are stimulated by the loudspeaker signals $x_i(k)$ described by vectors

$$\boldsymbol{x}_i(k) = [x_i(k), \cdots, x_i(k - L + 1)]^T, \quad (3)$$

and, then, the resulting echo estimates

$$\hat{y}_i(k) = \boldsymbol{w}_i^T(k)\boldsymbol{x}_i(k) \quad (4)$$

are subtracted from the microphone signal $y(k)$ to cancel the echoes.

The specific 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 a common sound source in the transmission room (e.g. a TV studio, *Fig.* 1). A straightforward extension of the mono AEC scheme using

$$\boldsymbol{w}(k) := [\boldsymbol{w}_1^T(k), \cdots, \boldsymbol{w}_P^T(k)]^T, \quad (5)$$

$$\boldsymbol{x}(k) := [\boldsymbol{x}_1^T(k), \cdots, \boldsymbol{x}_P^T(k)]^T \quad (6)$$

thus leads to very slow convergence of the adaptive filter towards the physically true echo paths [1, 2]. If the relation between the signals $x_i(k)$ is strictly linear, then there is a fundamental problem of non-uniqueness in the multi-channel case as was shown in [2]. In general, convergence to the true echo paths is necessary, since otherwise the AEC not only would have to track changes of the echo paths in the receiving room but also any changes of the crosscorrelation between the channels of the incoming audio signal, leading to sudden degradation of the echo cancellation performance [2]. The problem can be softened by some inaudible preprocessing of the loudspeaker signals for partial decorrelation of the channels, but sophisticated adaptation algorithms taking the cross-correlation into account are still necessary for M-C AEC [1].

Moreover, in a real-life environment, there are some more disturbances to the speech signal, apart from the interfering loudspeaker signals. The reverberation of the speech signal, background noise and/or other speakers make large-vocabulary ASR without any body-worn gear such as a headset still difficult. An effective approach to partly overcome these problems is to replace the personal microphone by a microphone array directing a beam

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

of increased sensitivity at the active talker. For the echo canceller however, this scenario presents a MIMO (multiple input and multiple output) system identification problem with P inputs (loudspeakers) and Q outputs (microphones).

2. COMBINING M-C AEC WITH BEAMFORMING MICROPHONE ARRAY

In [4] different strategies for combining AEC with a beamforming microphone array are discussed. Here we focus on the multichannel case of AEC in conjunction with a steerable microphone beam. Thereby we assume a filter and sum beamformer with arbitrary filters $a_q(k)$, $q = 1, \ldots, Q$ for the microphone signals. All the known fixed and adaptive beamforming approaches can be derived from this structure. As noted in [4], the two generic concepts for combining AEC and BF are either applying AEC separetely for each microphone signal or placing only one AEC behind the BF. The first concept obviously does not structurally differ from the single-microphone case in terms of AEC performance, however the second concept (*Fig.* 2) promises great computational savings. Apart from the general tracking problems of the AEC, which would become considerable in a straightforward implementation of this structure with adaptive BF [4], does it introduce any additional problems with M-C AEC? If no local speaker or background

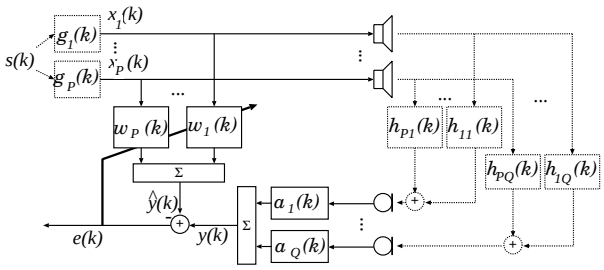

Fig. 2. Conventional M-C AEC structure and BF

noise is present, the error signal in *Fig.* 2 is

$$\begin{aligned} e(k) &= y(k) - \hat{y}(k) \\ &= \sum_{q=1}^{Q} a_q(k) * \left[\sum_{p=1}^{P} h_{pq}(k) * x_p(k) \right] - \\ &\quad - \sum_{p=1}^{P} w_p(k) * x_p(k). \end{aligned} \tag{7}$$

Assuming that the loudspeaker signals are obtained from a common source $s(k)$ via different acoustic paths $g_p(k)$ in the sending room, we have

$$x_p(k) = g_p(k) * s(k). \tag{8}$$

Together with Eq. (7), this finally leads to

$$\begin{aligned} e(k) &= s(k) * \left[\sum_{p=1}^{P} \left(\left(\sum_{q=1}^{Q} a_q(k) * h_{pq}(k) \right) - \right. \right. \\ &\quad \left. \left. - w_p(k) \right) * g_p(k) \right]. \end{aligned} \tag{9}$$

Thus, we see that for minimizing $E[|e(k)|^2]$ independently of $s(k)$ and the sending room (i.e. $g_p(k)$), the weights $w_p(k)$ of the AEC ideally have to model FIR systems according to

$$w_p(k) = \sum_{q=1}^{Q} a_q(k) * h_{pq}(k). \tag{10}$$

If the coefficients $a_q(k)$ are time-invariant, then Eq. (10) replaces the goal of identifying the true room impulse responses [2] for the system in *Fig.* 2.

In order to circumvent time-variant BF in the echo path of the AEC, decomposition of the BF into time-invariant and time-variant stages have been proposed [4]. 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. A *time-variant* weighted sum of the beam signals (voting) [3] is then moved *behind* the echo cancellation. An additional advantage of the structure is that external information of the speaker position via audio, video or multimodal object localization can be easily incorporated.

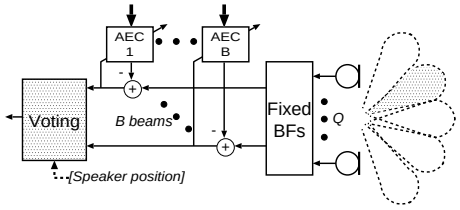

Fig. 3. AEC integrated into cascaded beamforming

The reduction of the number of M-C AECs from Q to B may be significant. However, considering the fact that most algorithms for M-C AEC are already very complex with only one microphone, identifying the electro-acoustic $P \times B$ MIMO system, described by $P \cdot B$ impulse responses Eq. (10) seems to be a computationally very demanding task.

In the following section we show how an efficient solution to this problem becomes possible for real-time systems.

3. COMPUTATIONALLY EFFICIENT M-C AEC FOR THE BEAM SIGNALS

3.1. A Separation Approach for MIMO System Identification

The total complexity of the AEC can be greatly reduced without detrimental effects on their convergence behaviour by decomposing the M-C AEC algorithms into parts that can be shared among all beams and other parts that have to be computed separately for each beam. The considerations are of course also applicable to the structure in which AEC is applied to each of the Q original microphone signals.

The recursive least squares (RLS) algorithm is regarded as upper reference for convergence speed and other adaptive algorithms can be seen as an approximation to it [5]. Despite of its high computational complexity, it has been also considered for M-C AEC due to the ill-conditioned nature of the problem [1]. The classical form of the RLS coefficient update can be written as

$$\boldsymbol{k}(k) = \boldsymbol{R}_{xx}^{-1}(k)\boldsymbol{x}(k), \tag{11}$$

$$\boldsymbol{w}(k+1) = \boldsymbol{w}(k) + \boldsymbol{k}(k)e^*(k). \tag{12}$$

It is known to be computationally very demanding due to the matrix inversion for the *Kalman gain vector* in Eq. (11) with a complexity of $O((P \cdot L)^2)$. From Eq. (11) we see that this main part only depends on the input data, not on the microphone or beam signals. Once this vector is known, the beam-dependent part Eq. (12) corresponds to an LMS algorithm. In the time domain the decomposed AEC in *Fig.* 4 consists of one Kalman gain calculation and B LMS algorithms. The Kalman gain can be efficiently calculated by *Fast RLS* algorithms (prediction part), which reduce the complexity of the RLS algorithm to $a \cdot L$, where $a \geq 7$ depends on the FRLS realization, but is $O(P^2)$. Similar considerations can be made for other adaptive algorithms (e.g. FNTF, FAP) as well [5].

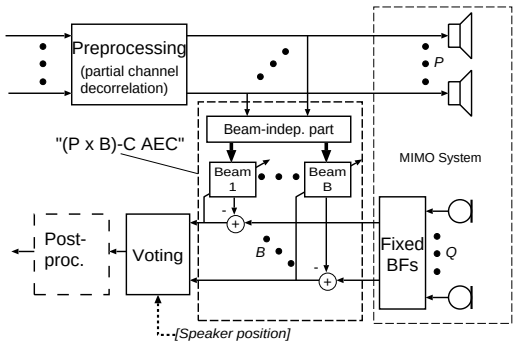

Fig. 4. Proposed Structure

The *relative computational savings* due to separation in case of the FRLS algorithm are

$$\begin{aligned} \zeta &= 1 - \frac{B \cdot 2PL + (a-2P)L}{B \cdot aL} \\ &= \left(1 - \frac{1}{B}\right)\left(1 - \frac{2P}{a}\right). \end{aligned} \quad (13)$$

As $a = O(P^2)$, we see that $0 < \zeta < 1$ increases along with the number of beams B and loudspeakers P. It is interesting to note that ζ is *independent* of L and therefore the same for the application in frequency subbands[1].

Recently, frequency-domain adaptive filtering has been introduced for M-C AEC [6] whose complexity is even much lower than that of the LMS algorithm and good performance comparing well with fast versions of high-order affine projection algorithms (FAP) can be obtained. In some cases it is even rivaling the RLS performance [1]. Note that obviously, the LMS complexity in the time-domain cannot be further reduced by the described separation, however simple LMS adaptation only is not suitable for M-C AEC.

3.2. Application to Frequency-Domain Adaptive Filtering

In contrast to the LMS, frequency-domain adaptive filtering allows clear savings by the separation approach despite of the already very low complexity of this technique. In this section we regard the unconstrained two-channel version following [6, 1] as an example. The algorithm is summarized in Table 1 in a separated form for B beams. The impulse responses consist of K partitions of length N each. δ_i denote regularization parameters, α is an overlap factor [6].

[1]Analysis of $x_i(k)$ by filter banks is also beam-independent, which leads to additional relative savings.

	Definitions
	$i, j = 1, 2$ (number of loudspeaker channel)
	$b = 1, \ldots, B$ (number of beam)
	$k = 0, \ldots, K-1$ (number of partition)
	$\boldsymbol{W} = diag\{[\mathbf{0}_{\mathbf{1}\times\boldsymbol{N}}\,\mathbf{1}_{\mathbf{1}\times\boldsymbol{N}}]\}$
	$\gamma = 1 - \beta$
	Beam-independent part
	Loudspeaker signals:
1	$\boldsymbol{X}_i(m) = diag\{\boldsymbol{F}[x_i(m\frac{N}{\alpha} - N + 1) \cdots x_i(m\frac{N}{\alpha} + N)]^T\}$
	Power spectrum estimation:
2	$\boldsymbol{S}_{i,j}(m) = \beta \boldsymbol{S}_{i,j}(m-1) + \gamma \boldsymbol{X}_i^*(m)\boldsymbol{X}_j(m)$, where $\boldsymbol{S}_{j,i}(\cdot) = \boldsymbol{S}_{i,j}^*(\cdot)$
3	$\tilde{\boldsymbol{S}}_{i,i}(m) = \boldsymbol{S}_{i,i}(m) + diag\{[\delta_0 \cdots \delta_{2N-1}]\}$
4	$\boldsymbol{T}(m) = [\boldsymbol{I}_{\boldsymbol{2N}\times\boldsymbol{2N}} - -\rho^2 \boldsymbol{S}_{1,2}^*(m)\boldsymbol{S}_{1,2}(m)\{\tilde{\boldsymbol{S}}_{1,1}(m)\tilde{\boldsymbol{S}}_{2,2}(m)\}^{-1}]$
5	$\boldsymbol{S}_i(m) = \tilde{\boldsymbol{S}}_{i,i}(m)\boldsymbol{T}(m)$
	Beam-independent weights for filter updates:
6	$\boldsymbol{\Delta}_{1,k}(m) = \boldsymbol{S}_1^{-1}[\boldsymbol{X}_1^*(m-k) - \rho \boldsymbol{S}_{1,2}\tilde{\boldsymbol{S}}_{2,2}^{-1}\boldsymbol{X}_2^*(m-k)]$
7	$\boldsymbol{\Delta}_{2,k}(m) = \boldsymbol{S}_2^{-1}[\boldsymbol{X}_2^*(m-k) - \rho \boldsymbol{S}_{2,1}\tilde{\boldsymbol{S}}_{1,1}^{-1}\boldsymbol{X}_1^*(m-k)]$
	Beam-dependent part
8	$\hat{\boldsymbol{Y}}_{i,b}(m) = \sum_{k=0}^{K-1} \boldsymbol{X}_i(m-k)\hat{\boldsymbol{H}}_{i,k,b}(m)$
9	$\tilde{\boldsymbol{e}}_b(m) = [\mathbf{0}_{\mathbf{1}\times\boldsymbol{N}} y_b(m\frac{N}{\alpha} + 1) \cdots y_b(m\frac{N}{\alpha} + N)]^T - -\boldsymbol{W}\boldsymbol{F}^{-1}[\hat{\boldsymbol{Y}}_{1,b}(m) + \hat{\boldsymbol{Y}}_{2,b}(m)]$
10	$\tilde{\boldsymbol{E}}_b(m) = \boldsymbol{F}\tilde{\boldsymbol{e}}_b(m)$
11	$\hat{\boldsymbol{H}}_{i,k,b}(m+1) = \hat{\boldsymbol{H}}_{i,k,b}(m) + \mu_b \Delta_{i,k}(m)\tilde{\boldsymbol{E}}_b(m)$
12	$\boldsymbol{e}_b(m) =$last $\frac{N}{\alpha}$ elements of $\tilde{\boldsymbol{e}}_b(m)$

Table 1. $2 \times B$-channel frequency-domain adaptive filtering

The complexity of this algorithm is clearly dominated by FFT calculations. The number of real multiplications per output sample is $\alpha[2\log_2(2N) + 24K + 50 + 4/N] + \alpha \cdot B[2\log_2(2N) + 24K - 6 + 4/N]$ and the number of real additions is $\alpha[6\log_2(2N) + 16K + 14 + 8/N] + \alpha \cdot B[6\log_2(2N) + 16K - 8 + 8/N]$, assuming that FFT is carried out by the split radix algorithm and several quantities in Tab. 1 are real for real data. *Fig.* 5 illustrates the total number of floating point operations (i.e. sum of real MULs and ADDs) per output sample for FRLS and LMS in the time domain (a) and the mentioned approach in the frequency domain (b) for $\alpha = 4$. For all curves, $P = 2$ ($a = 28$ for two-channel FRLS [1]) and $L = 4096$ is assumed. Solid lines stand for the separated, dashed lines for the original approach. The same relations between the curves as in (a) are also valid for the FRLS in frequency subbands. Let us assume e.g. $B = 7$, $N = 1024$, $K = 4$ (i.e. $L = 4096$), $\alpha = 4$. With these parameters, the described algorithm (separated, 50% additional savings) is for 7 beams about 4 times as efficient as the LMS algorithm with a single microphone and thus provides also enough scope for necessary double-talk detection for the AEC. The entire system can be realized on a regular Intel-based computer (1GHz, Linux) with a multi-channel soundcard at a sampling frequency of 12kHz.

4. EVALUATION WITH ASR

4.1. Setup for Evaluation

In order to verify the effectiveness of the combined approach, we apply our system to a large-vocabulary continuous ASR (dictation system *Dragon Naturally Speaking*). The ASR was trained with our system in the actual receiving room. Texts for evaluation and for speaker/setup adaptation were non-overlapping, however, the

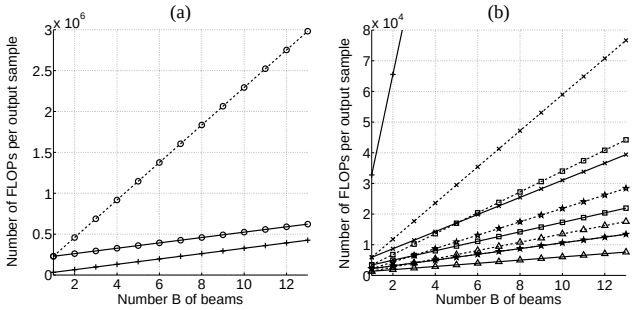

Fig. 5. Computational Complexity for Separated (solid), Orig. (dashed) Approach with FRLS ($\circ$), LMS ($+$), and Table 1 with Different Block Lengths N: 256 ($\times$), 512 ($\square$), 1024 ($\star$), 4096 ($\triangle$)

vocabulary of the dictated text for the evaluation was in an already known context. For evaluations relevant to real-life situations this is meaningful for large vocabulary dictation systems. *Fig.* 6 shows the setup for the evaluation, which was carried out in two different environments (length $l\times$ width $w\times$ height): an anechoic chamber (254cm $\times$ 274cm $\times$ 238cm) and in a real office room (480cm $\times$ 350cm $\times$ 350cm) with reverberation times of $60msec$ and $300msec$, resp. For the fixed beamforming, a nested line array of 11 first order differential microphones and a delay-and-sum beamformer was used. The spacing for the lower frequency range ($240Hz - 2.8kHz$) is $8cm$ and for the higher frequency range ($2.8 - 5.5kHz$), it is $4cm$. Loudspeakers and microphones are located at the same height. For the loudspeaker signals, we chose the solo music *prelude in c major*(piano). Solo music with a spatially fixed instrument is known to be demanding for M-C AEC, however the two channels were separately encoded using a psychoacoustic model, which reduces the cross-correlation [1]. M-C AEC adaptation was carried out as described in sect. 3.2 with a stepsize control based on [7]. The distance between speaker and microphone was $2m$. Speaker and loudspeaker signals had the same level; the local interference (white noise) level was $20dB$ below the speaker level.

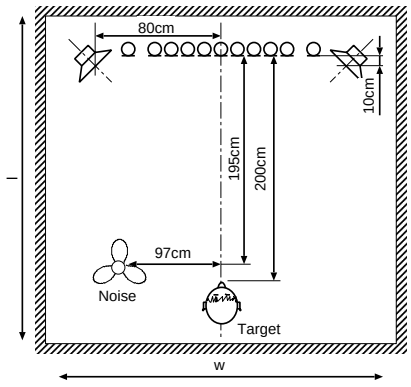

Fig. 6. Setup for Evaluation

Interference	Single Mic.	BF	Single Mic.+AEC	BF +AEC
no interf.	84%	91%	84%	91%
local interf.	65%	86%	65%	86%
local interf+echoes	48%	61%	63%	85%

Table 2. Word Recognition Accuracies in the Anechoic Chamber

Interference	Single Mic.	BF	Single Mic.+AEC	BF +AEC
no interf.	62%	84%	62%	82%
local interf.	58%	81%	58%	80%
local interf+echoes	36%	50%	56%	80%

Table 3. Word Recognition Accuracies in the Office Room

4.2. Recognition Results

Tables 2 and 3 summarize the measured word recognition accuracies obtained in the respective environment. Note that without echoes the AEC adaptation was halted.

5. CONCLUSIONS

In this paper, a computationally efficient concept and an example system for M-C AEC in conjuntion with a steerable beamforming microphone array has been presented. It was compared in terms of complexity and word accuracies of a large vocabulary speech recognizer, which are considerably improved by the system compared to a single microphone or BF/AEC only. By further processing the input of the baseline ASR (residual disturbances), additional gains can be expected.

6. 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, August 1995, pp. 148-151.

[3] W. Kellermann, "A self-steering digital microphone array," Proc. ICASSP 1991, pp. 3581-3584.

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

[5] G. Glentis, K. Berberidis and S. Theodoridis, "Efficient Least Squares Adaptive Algorithms For FIR Transversal Filtering - A Unified View," *IEEE SP Mag.*, pp. 13-41, July 1999.

[6] 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.

[7] S. Yamamoto and S. Kitayama, "An Adaptive Echo Canceller With Variable Step Gain Method," Trans. of the IECE of Japan, vol. E 65, no. 1, pp. 1-8, 1982.