INTERPOLATION OF MVDR BEAMFORMER COEFFICIENTS FOR JOINT ECHO CANCELLATION AND NOISE REDUCTION

*Markus Kallinger *, Joerg Bitzer **, Karl-Dirk Kammeyer **

* University of Bremen, FB 1, Dept. of Communications Engineering
P.O. Box 330 440, D–28334 Bremen, Germany
email: kallinger@comm.uni-bremen.de
** Houpert Digital Audio,
Anne-Conway-Str. 1, D–28359 Bremen, Germany,
e-mail: j.bitzer@hda.de

ABSTRACT

This contribution deals with a new multi-microphone algorithm for joint acoustic echo cancellation (AEC) and noise reduction (NR). It is known that microphone arrays are useful to suppress the far-end speaker signal, if the AEC is not converged. Solutions based on adaptive beamformers are especially suited to adjust themselves to the noise field generated by the loudspeaker of a hands-free system. However, they have some problems with signal cancellation and potential interaction between the AEC and the beamformer which leads to a poor performance of the AEC. We propose a new approach with a partly adaptive beamformer which can be modified within two special cases: A superdirective beamformer and an adaptive beamformer which only becomes active as long as the AEC is diverged, e.g. after a modification of the echo path impulse response. We will introduce a criterion based on the convergence of the AEC to steer the switching between both beamformers. In order to avoid a sudden exchange of the beamformer coefficients, the new scheme interpolates between 'old' and 'new' coefficients and enables a smooth crossover by this.

1. INTRODUCTION

In hands-free speech communication, we have to face the following fundamental problems:

- Acoustic echoes introduced by the loudspeaker of the hands-free system.
- Additive disturbances with unknown statistics and spatial coherence.
- Linear distortion of the desired signal by reverberation.

The ideal solution to suppress acoustic echoes is the AEC. Depending on the acoustic environment, the adaptive filter has to be very long. This results into slow convergence of the AEC. In order to support the AEC during initial convergence, an adaptive beamformer can be used. It provides additional echo attenuation as it is capable to adapt itself to the noise field which is generated by the loudspeaker of the hands-free system and the environmental noise sources. However, an adaptive beamformer requires a reliable voice activity detection (VAD) to prevent signal cancellation [1]. Furthermore, using an adaptive beamformer does not allow arbitrary setups with an AEC, since rapidly changing filters heavily disrupt the AEC's convergence, if the beamformer is in the front. Several ways have been proposed to avoid this problem [2, 3].

In order to address the problem of noise reduction in enclosures, we can use a superdirective beamformer designed for a spherically isotropic noise field, since the noise-field in reverberant, mid-sized rooms tends to be diffuse [4]. Additionally, the beamformer has some capabilities for dereverberation of the desired speech signal [5].

Therefore, we have two different beamformers which are good choices for different specific situations: An adaptive beamformer which provides echo attenuation as long as the AEC is not converged, or a superdirective beamformer to reduce reverberation and diffuse noise when the AEC works properly. Hence, we suggest to switch between these two limiting cases depending on the adaptation status of the AEC. This is a robust solution for a simple and efficient beamformer design which leads to a high quality of speech.

In order to get the status of the AEC's adaptation we estimate the **E**cho **R**eturn **L**oss **E**nhancement (ERLE). This can be done by using two estimation methods which are well known within different contexts: The *minimum statistics estimation method* that was proposed for robust spectral estimation [6] and the *delay coefficients method* which was originally used for the step-size control of an AEC [7]. The suggested ERLE-estimation is described in section 2.

Since a 'hard switching' of the beamformer coefficients might introduce some undesired artifacts, we propose a sliding linear interpolation of the coefficients of the two limiting beamformer designs which both fulfill the **M**inimum **V**ariance **D**istortionless **R**esponse design rule (MVDR) [8]. In section 3, we will show that a linear interpolation of two beamformers' coefficients does not corrupt the constraint of the undistorted look direction. A cross-fade of the coefficients results in a beamformer with a mixed spatial and spectral behaviour.

Section 4 shortly illustrates the general conditions for the simulations with the new combined structure. Some results will be given as well. In section 5, we conclude the paper.

2. ERLE ESTIMATION

Our proposal to switch the beamformer coefficients depending on the adaptation status of the AEC is based on the estimation of the ERLE. When the ERLE is low we switch to the adaptive beamformer in order to achieve additional echo attenuation. As long as the ERLE suffices and the AEC works properly the combined system employs a superdirective beamformer. Unfortunately the ERLE is not accessible separately; it is defined by $E\{d^2(k)\}/E\{\epsilon^2(k)\}$, where $\epsilon(k) = d(k) - \hat{d}(k)$ is the residual echo and $E\{\cdot\}$ is the expectation operator. The signals are illustrated in figure 1. The filter that models the **R**oom **I**mpulse **R**esponse (RIR) is included in the time-variant coefficient vector $\mathbf{h}_0(k)$. The compensation filter in the AEC is described by the coefficient vector $\hat{\mathbf{h}}_0(k)$. We introduce a delay of $P = 40$ samples into the echo path to apply the delay coefficients method. We define the system mismatch vector as

$$\mathbf{m}_0(k) = \mathbf{h}_0(k) - \hat{\mathbf{h}}_0(k)\,. \tag{1}$$

With the far-end speech signal $x(k)$ and assuming a white excitation signal we get

$$E\{\epsilon^2(k)\} = ||\mathbf{m}_0(k)||^2 E\{x^2(k)\}\,. \tag{2}$$

The system mismatch $||\mathbf{m}_0(k)||^2$ can be estimated, since we know that the first $P = 40$ samples in $\mathbf{h}_0(k)$ have to be zero [9]:

$$||\mathbf{m}_0(k)||^2 \approx \overline{p_{m_0}}(k) = \frac{Q+P}{P}\sum_{i=0}^{P-1}\hat{h}^2_{0,i}(k)\,. \tag{3}$$

Thus, the power of the residual echo can be estimated by $E\{\epsilon^2(k)\} \approx \overline{p_{m_0}}(k)E\{x^2(k)\}$. Ambient noise $n(k)$ and – in periods of double talk – near-end speech $s(k)$ will interfere with the echo signal $d(k)$. So, we cannot simply use the microphone signal $y(k) = s(k) + n(k) + d(k)$ to estimate the ERLE. Again, we suppose a white excitation signal $x(k)$. Therefore, the energy of the echo path impulse response is $||\mathbf{h}_0(k)||^2 = E\{d^2(k)\}/E\{x^2(k)\}$. We assume that the interfering signals $n(k)$ and $s(k)$ are not correlated with $x(k)$. Hence, an estimated value for $||\mathbf{h}_0(k)||^2$ will rise in periods of double talk, since we must use the microphone signal $y(k)$ instead of the echo signal $d(k)$. If we suppose a negligible ambient noise level, the lapse of the estimation of $||\mathbf{h}_0(k)||^2$, $\overline{p_{h_0}}(k)$, will show short peaks in periods of double talk due to the additive 'disturbing' near-end speech signal $s(k)$. Whereas the desired quotient

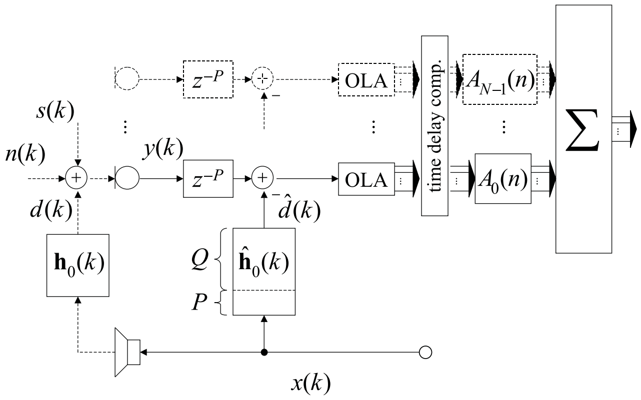

Figure 1: Signal model of a combined structure with a multi-channel AEC-unit in front of a beamformer.

$E\{d^2(k)\}/E\{x^2(k)\}$ changes slowly which presumes that the echo path alters slowly as well. So, the influence of $s(k)$ on $\overline{p_{h_0}}(k)$ can be suppressed by using the minimum statistics estimation method [6] on $\overline{p_{h_0}}(k)$ which is well suited to suppress short rising peaks. In our algorithm the 'memory window' was set to a length of $2s$ which suffices to suppress even long and rather stationary periods of near-end speech. The window is updated every $0.4s$. Finally, the estimations of the powers of the echo signal and the residual echo are

$$\begin{aligned}\overline{d^2}(k) &= \overline{p_{h_0}}(k)\overline{x^2}(k) \quad \text{and} & (4)\\ \overline{\epsilon^2}(k) &= \overline{p_{m_0}}(k)\overline{x^2}(k)\,. & (5)\end{aligned}$$

The quotient $\overline{d^2}(k)/\overline{\epsilon^2}(k)$ leads to the estimated ERLE. Note that we compute all signal powers by a first-order IIR filter

$$E\{x^2(k)\} \approx \overline{x^2}(k) = (1-\gamma)x^2(k) + \gamma\overline{x^2}(k-1)\,. \tag{6}$$

The factor γ was set to correspond to a time constant of $40ms$. Figure 2 shows the lapse of the estimated ERLE compared to the actual ERLE measured in our simulation environment. We have simulated a RIR with a distance of $1m$ between the far-end speaker's loudspeaker and the microphone. The reverberation time τ_{60} was set to $100ms$ (we employed the image method by Allen and Berkley [10]). In order to enable a quick adaptation of the AEC (512 adaptive coefficients), we used an *Affine Projection Algorithm* (APA, [11]) with a projection order of 4. We switched on a step-size control as proposed in [7] after 70,000 samples to obtain robustness against double talk.

The RIR was modified after 55,000 samples. The intermediate decay of the ERLE can be seen clearly in both plots in figure 2. The far-end speaker pauses at 40,000 and 90,000 samples which results in short collapses of the actually measured ERLE. The estimation of the ERLE is frozen when the far-end speech signal power becomes too low. The near-end speaker, who is active between sample 80,000 and 100,000, does not disturb the ERLE estimation notedly. Although the estimation does not follow the actual ERLE curve precisely, it enables a reliable estimation of the AEC's adaptation status.

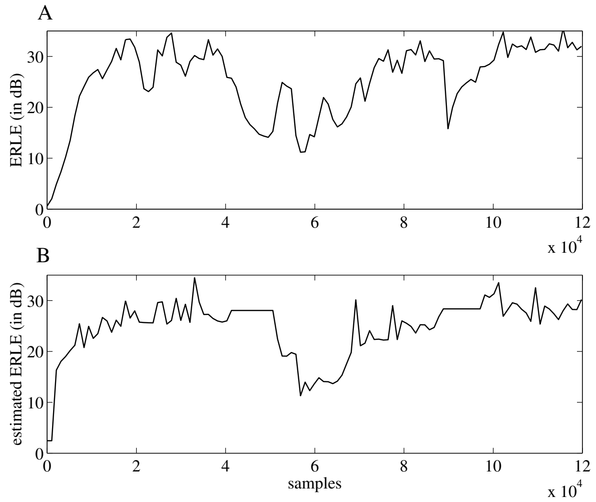

Figure 2: ERLE (A) and estimated ERLE (B) measured at an AEC. There is a sudden modification of the RIR at 55,000 samples. A near-end speaker is active between sample 80,000 and 100,000. The far-end speaker pauses at 40,000 and 90,000 samples.

3. COEFFICIENT INTERPOLATION

As we could see in section 2, the proposed method for estimating the ERLE delivers results which allow a switching of beamformer coefficients depending on the adaptation status of the AEC. According to figure 2, for example, we would invent a threshold of $20dB$ for the estimated ERLE. Below the threshold we switch on the adaptive beamformer and above of it we stop the adaptation and switch to the superdirective coefficients. Since the beamformer as well as the time delay compensation (see figure 1) work in the frequency domain, we could exchange the coefficients abruptly between two succeeding signal blocks. However, depending on the speech signal this might result into undesired artifacts. Instead, we interpolate the complex beamformer coefficients using a first-order IIR filter as shown in equation (6) at each discrete frequency index n. In the following, we point out that the constraint of a distortionless look direction within the MVDR design rule is not corrupted by this.

Let $\mathbf{A}(n)$ be the $N \times 1$ frequency dependent vector

$$\mathbf{A}(n) = [A_0(n), \cdots, A_{N-1}(n)]^T \tag{7}$$

of a beamformer's coefficients. $A_i(n)$ is the n-th complex filter coefficient in the frequency domain of the i-th microphone channel. N is the number of microphones. We consider two separately designed coefficient vectors:

- $\mathbf{A^{SD}}(n)$: Superdirective beamformer.
- $\mathbf{A^{opt}}(n)$: Adaptive MVDR beamformer in an open-loop architecture with continuous estimation of the cross power spectrum densities between the microphone signals.

An interpolated beamformer possesses the coefficient vector $\mathbf{A^{int}}(n)$. The new vector has to fulfill the constraint for a distortionless look direction

$$\sum_{i=0}^{N-1} A_i^{int}(n) \stackrel{!}{=} 1 \qquad \forall\, n. \tag{8}$$

We perform the coefficient interpolation according to

$$\mathbf{A^{int}}(n) = (1-\delta)\mathbf{A^{opt}}(n) + \delta \mathbf{A^{SD}}(n) \tag{9}$$

with $0 \le \delta \le 1$. By inserting the rows of (9) into (8), we can see that any linear combination of MVDR beamformers leads to a beamformer with a distortionless response in the look direction:

$$\begin{aligned} \sum_{i=0}^{N-1} A_i^{int}(n) &= (1-\delta) \underbrace{\sum_{i=0}^{N-1} A_i^{opt}(n)}_{1} + \delta \underbrace{\sum_{i=0}^{N-1} A_i^{SD}(n)}_{1} \\ &= 1 \qquad \forall\, n. \end{aligned} \tag{10}$$

Beam pattern **C** in figure 3 illustrates the behaviour of a new interpolated beamformer. We employed a linear array of four microphones in endfire steering to π with a spacing of $5cm$ at a sampling frequency of $8kHz$. The properties of the interpolated beamformer seem to result from the two primary beamformers.

4. SIMULATION RESULTS

In the proposed combined system we employ separate AECs for each of the four microphone channels in front of the beamformer (The dashed structure in fig. 1 indicates the additional AECs). Each AEC is based on the same reference channel. The *fast fourier transformation* (FFT) for a fast convolution in the frequency domain is executed by an extended *overlap-add* structure (OLA) for filtering with time-variant, non-causal impulse responses. The FFT-length accounts to 512 with a hopsize of 128. For the simulations, we have used AECs with AP algorithms as described in section 2.

Plot 4.**A** shows the ERLEs which were measured at the beamformer-unit. There is no double talk in the selected period but the four RIRs between the loudspeaker and each microphone change abruptly at 60,000 samples. Since the estimated ERLE falls below the chosen threshold of $20dB$ (see plot 4.**B**), the adaptive beamformer is turned on and the superdirective coefficients are interpolated towards the adaptive coefficients. In these periods with misaligned AECs, we can see that the adaptive beamformer (dashed line in plot 4.**A**) delivers about $5dB$ of additional ERLE compared to the superdirective beamformer (solid line). The adaptive beamformer supports the AECs at the beginning and after the modification of the RIRs.

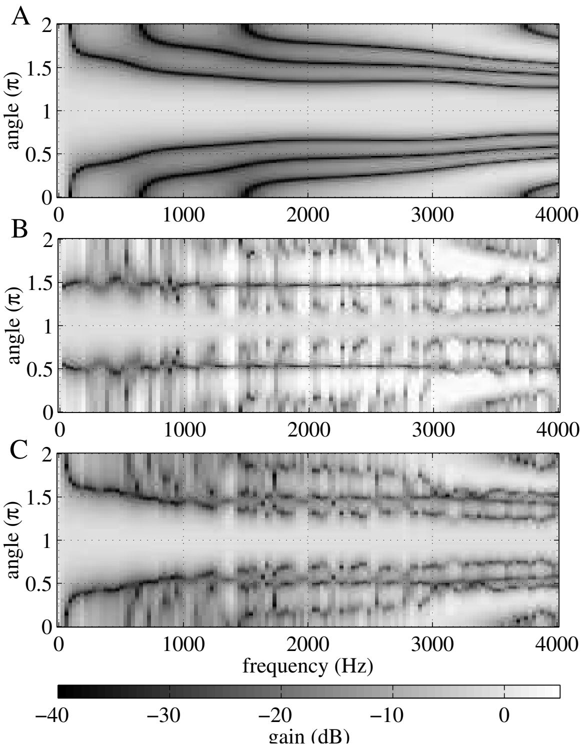

Figure 3: Beam pattern of a superdirective beamformer (A), an optimally designed beamformer (B), and linear interpolation between them with $\delta = 0.75$ (C).

5. CONCLUSIONS

In this contribution, we have proposed a simple and effective combined structure with a multi-channel AEC-unit in front of a partly adaptive beamformer which is controlled by the AECs' adaptation status. In order to gather the adaptation status, we have introduced a reliable way to estimate the ERLE. In the system, an adaptive beamformer can support the AEC-unit in periods of misalignment. Furthermore, we can exploit the benefits of a superdirective beamformer, when the AEC-unit works properly.

6. REFERENCES

[1] D. Compernolle, W. Ma, F. Xie, and M. Diest, "Speech Recognition in Noisy Environments with the Aid of Microphone Arrays," *EURASIP Speech Communication*, vol. 9, pp. 433–442, 1990.

[2] W. Kellermann, "Strategies for Combining Acoustic Echo Cancellation and Adaptive Beamforming Microphone Arrays," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, vol. 1, (Munich, Germany), pp. 219–222, April 1997.

[3] W. Herbordt and W. Kellermann, "GSAEC – Acoustic Echo Cancellation embedded into the Generalized Sidelobe Canceller," in *Proc. European Signal Processing Conference (EUSIPCO)*, (Tampere, Finland), September 2000.

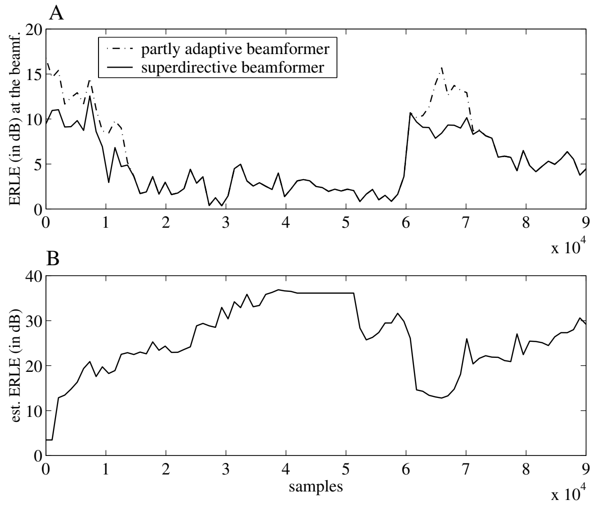

Figure 4: ERLE at the beamformer after a preceeding multichannel AEC (A). The RIRs are exchanged at 60,000 samples. The adaptive beamformer is switched on, when the estimated ERLE (B) falls below $20dB$.

[4] T. J. Schultz, "Diffusion in reverberant rooms," *J. Sound and Vibration*, vol. 16, no. 1, pp. 17–28, 1971.

[5] G. W. Elko, "Superdirectional microphone arrays," in *Acoustic Signal Processing for Telecommunication* (S. L. Gay and J. Benesty, eds.), ch. 10, pp. 181–235, Kluwer Academic Publishers, 2000.

[6] R. Martin, "Spectral Subtraction Based on Minimum Statistics," in *European Signal Processing Conference (EUSIPCO–94)*, (Edinburgh, UK), pp. 1182–1185, September 1994.

[7] C. Breining, P. Dreiseitel, E. Hänsler, A. Mader, B. Nitsch, H. Puder, T. Schertler, G. Schmidt, and J. Tilp, "Acoustic Echo Control – An Application of Very-High-Order Adaptive Filters," *IEEE Signal Processing Magazine*, pp. 42–69, July 1999.

[8] S. Haykin, *Adaptive Filter Theory*. Prentice Hall, 1996.

[9] A. Mader, H. Puder, and G. Schmidt, "Step-Size Control for Acousatic Echo Cancellation Filters – an Overview," *Elsevier Signal Processing*, vol. 80, pp. 1697–1719, September 2000.

[10] J. B. Allen and D. A. Berkley, "Image Method for Efficiently Simulating Small–Room Acoustics," *J. Acoust. Soc. Amer.*, vol. 65, pp. 943–950, 1979.

[11] K. Ozeki and T. Umeda, "An Adaptive Filtering Algorithm Using Orthonormal Projection to an Affine Subspace and it's Properties," *Electronic and Communications in Japan*, vol. 67-A, pp. 126–132, Feb 1984.