MULTI-MICROPHONE RESIDUAL ECHO ESTIMATION

Markus Kallinger, Karl-Dirk Kammeyer

University of Bremen, FB 1
Dept. of Communications Engineering
P.O. Box 330 440
D–28334 Bremen, Germany
kallinger@ant.uni-bremen.de

Jörg Bitzer

Houpert Digital Audio
Anne-Conway-Str. 1
D–28359 Bremen, Germany

j.bitzer@hda.de

ABSTRACT

Post-filters are a powerful extension to improve echo attenuation when combined with the well-known echo canceller. In order to guarantee high quality of the transmitted speech signal the primary purpose of a post-filtering system is to estimate the power spectral density (PSD) of the residual echo at the output of the echo canceller as accurately as possible. In this contribution, we introduce a novel technique to estimate the residual echo by using a microphone array. The robustness against double-talk and other additive interferences is reached by means of minimum statistics and further enhanced by exploiting spatial information. Simulation results show that the new methods are able to estimate the residual echo even under adverse conditions.

1. INTRODUCTION

In recent proposals for high-quality hands free systems, the combination of beamforming techniques and acoustic echo cancellers (AECs) has become more and more popular [1]. AECs are the optimum solution to avoid the acoustic feedback of the sent speech signal. However, in real-world applications, the performance of AECs are limited by additive interferences as well as time variant systems, which have to be identified [2]. Apart from their capabilities to enhance the near-end speech signal beamformers can support the AEC in terms of echo attenuation.

In an implemented system, the computational load roughly rises by the number of microphones in the beamforming array, if an AEC resides after each microphone. An alternative would be to position one AEC at the output of the beamformer. However, this involves disturbing influences of the beamformer onto the AEC, when the beamformer or a preceding steering unit is changing fast. One solution to this problem represents the constraint of the beamformer's steering unit to a fixed number of "discrete looking directions". In turn, it becomes necessary to have the same number of AECs running in parallel for each looking direction [1]. When a certain limit in matters of the computational power is reached, the order of the AECs' adaptive filters have to be shortened.

Post-filters, which are designed for the residual echo after the AEC, can enhance the echo attenuation. In addition, they represent a quickly converging, redundant unit to the AEC, which works in a different manner [3, 4]. In this paper we introduce a new post-filter for residual echo attenuation. This post-filter makes use of both information in the reference signal path and spatial information, which becomes available by the employment of a microphone array.

In the next section we investigate different ways to estimate the residual echo within a multi-microphone setup. Robustness against double-talk can be gathered according to section 3. All simulation results are given in section 4. In section 5 we summarize the basic statements of the paper 4.3.

2. ESTIMATING THE RESIDUAL ECHO

Compared with known multi-microphone post-filters for noise reduction [5], we can now exploit the advantage that a reference signal $X(m,l)$ (i.e. the far-end speech signal) is available. $X(m,l)$ is gathered with the help of the discrete Fourier transform (DFT) at a length of L_{DFT} from the signal $x(k)$. To unify the upcoming illustrations, all signals will be described in the frequency domain with a frame index l and a discrete frequency index m. Figure 1 shows our basic signal model, which employs a multi-microphone AEC with the compensation filters' transfer functions $\mathbf{C}_i(m,l)$, a fixed beamformer with the transfer functions $A_i(m,l)$, and a single-channel post-filter $P(m,l)$. The index i denotes the microphone channel and M is the number of microphones. To consider the system orders of the room impulse response (RIR) $\mathbf{H}_i(m,l)$, the AEC $\mathbf{C}_i(m,l)$, and the system misalignment $\mathbf{D}_i(m,l)$ we introduce the vectors

$$
\begin{aligned}
\mathbf{H}_i(m,l) &= \Big[H_{i,0}(m,l) \quad \cdots \quad H_{i,L'_H-1}(m,l)\Big], \quad (1)\\
\mathbf{C}_i(m,l) &= \Big[C_{i,0}(m,l) \quad \cdots \quad C_{i,L'_{AEC}-1}(m,l) \\
&\qquad 0 \quad \cdots \quad 0\big], \quad (2)\\
\mathbf{D}_i(m,l) &= \mathbf{H}_i(m,l) - \mathbf{C}_i(m,l), \quad (3)\\
\mathbf{X}(m,l) &= \big[X(m,l) \quad \cdots \quad X(m,l-L'_H+1)\big]^T. \quad (4)
\end{aligned}
$$

$L_H = L'_H L_{DFT}$ and $L_{AEC} = L'_{AEC} L_{DFT}$ are the lengths of the echo path impulse response and the AEC filter, respectively.

Each signal $Y_i(m,l)$ consists in the near-end signal $S_i(m,l)$ and the echo signal $\Psi_i(m,l) = \mathbf{H}_i(m,l)\mathbf{X}(m,l)$. Strictly speaking, a noise signal $N_i(m,l)$ should be considered here as well. However, internal simulations have shown, that the newly proposed system is robust against ambient noise up to a signal-to-noise ratio (SNR) of 20 dB. Therefore, any further noise signals are omitted in this paper. The signal to be estimated is the residual echo $\Xi_i(m,l) = \mathbf{D}_i(m,l)\mathbf{X}(m,l)$. The AECs' output signals $E_i(m,l)$ contain the residual echoes $\Xi_i(m,l)$ and the speech sig-

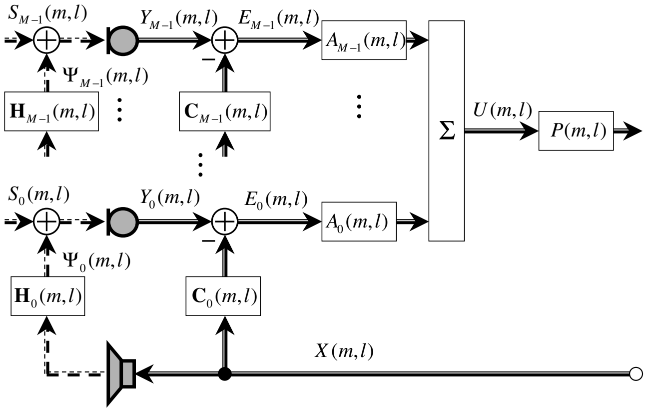

Fig. 1. Frequency domain signal model of acoustic echo cancellers in front of a beamformer with a succeeding post-filter.

nals $S_i(m,l)$. The residual echo at the beamformer's output is

$$\Xi_B(m,l) = \sum_{i=0}^{M-1} A_i(m,l)\Xi_i(m,l). \tag{5}$$

$U(m,l)$, the beamformer's output signal, results from $E_i(m,l)$ in the same way. Note that the steering of the microphone array (and thus, the compliance with the distortionless response condition [6]) is carried out by linear phase terms in the frequency domain. These terms are already implemented in the beamformer filters $A_i(m,l)$. We assume that the near-end signal at the beamformer output $S_B(m,l)$ can be reconstructed almost ideally and that the following relation holds for all microphones channels i:

$$S_i(m,l) \approx S_B(m,l) = \sum_{i=0}^{M-1} A_i(m,l)S_i(m,l). \tag{6}$$

Finally, we design a Wiener post-filter by the assumption of statistically independent signals $S_B(m,l)$ and $\Xi_B(m,l)$

$$P(m,l) = \frac{\Phi_{S_B S_B}(m,l)}{\Phi_{SS}(m,l) + \Phi_{\Xi_B\Xi_B}(m,l)}. \tag{7}$$

We obtain the estimated residual echoes $\hat{\Xi}_i(m,l)$ via estimates $\hat{\mathbf{D}}_i(m,l)$ of the system misalignment transfer functions $\mathbf{D}_i(m,l)$. $\hat{\mathbf{D}}_i(m,l)$ is a vector, which is defined in the same way as illustrated in equation (3). However, its length L'_{SME} should be smaller than L'_H for complexity reasons. Furthermore, we define

$$\hat{\mathbf{\Phi}}_{XX}(m,l) = \left[\hat{\Phi}_{XX}(m,l) \ \cdots \ \hat{\Phi}_{XX}(m,l-L'_{SME}+1)\right] \tag{8}$$
$$\hat{\mathbf{\Phi}}^{-1}_{XX}(m,l) = \left[\hat{\Phi}^{-1}_{XX}(m,l) \ \cdots \ \hat{\Phi}^{-1}_{XX}(m,l-L'_{SME}+1)\right].$$

$\hat{\mathbf{\Phi}}_{XE_i}(m,l)$ is defined in a similar way except for the difference that in its jth element, $E(m,l)$ is correlated with $X(m,l-j+1)$. All estimations of PSDs $\hat{\Phi}$ are calculated using Welch's method with recursive smoothing. Finally, we can set up the Wiener-Hopf equation in the frequency domain

$$\hat{\mathbf{D}}_i(m,l) = \hat{\mathbf{\Phi}}_{XE_i}(m,l) \otimes \hat{\mathbf{\Phi}}^{-1}_{XX}(m,l). \tag{9}$$

$\otimes$ denotes the element-by-element vector multiplication. An extended description of the computation of $\hat{\mathbf{D}}_i(m,l)$ can be found in [4]. In contrast to the single-channel solution, which is treated there, we can choose between three methods to compute the residual echo at the beamformer's output $\hat{\Xi}_B(m,l)$:

1. The first possibility is to calculate $\Xi_i(m,l)$ at each microphone channel and lead them through the beamformer as illustrated in equation (5). This option demands M estimators, which are based on the common reference signal $X(m,l)$.
2. Since for the application of a Wiener filter only the estimated PSD $\hat{\Phi}_{\Xi_B\Xi_B}(m,l)$ is required, it might suffice to calculate the mean of the PSDs of the residual echo signals in the microphone channels like

$$\hat{\Phi}_{\Xi_B\Xi_B}(m,l) = \frac{1}{M}\sum_{i=0}^{M-1} \hat{\Phi}_{\Xi_i\Xi_i}(m,l). \tag{10}$$

   This method involves a certain bias, because the beamformer usually provides some echo attenuation and this estimation of $\hat{\Phi}_{\Xi_B\Xi_B}(m,l)$ will be too large. On the other hand, the variance in each of the estimates $\hat{\Phi}_{\Xi_i\Xi_i}(m,l)$ could be reduced by computing the mean.
3. $\hat{\Phi}_{\Xi_B\Xi_B}(m,l)$ can be computed directly as well. This can be done, if we try to obtain the combined system

$$\mathbf{D}_B(m,l) = \sum_{i=0}^{M-1} \left(\mathbf{H}_i(m,l) - \mathbf{C}_i(m,l)\right) A_i(m,l). \tag{11}$$

   However, no multi-channel information can be utilized, because we have to replace $\hat{\Phi}_{XE_i}(m,l)$ by $\hat{\Phi}_{XU}(m,l)$ for a calculation of the system misalignment transfer function according to equation (9).

In section 4.1 we will compare these three new approaches on the basis of simulation results.

3. ROBUSTNESS AGAINST DOUBLE-TALK

In [4], we introduced a new technique to suppress interferences during the estimation of the system misalignment transfer function with the help of minimum statistics [7]. The basic steps of this procedure are outlined in the next section. In section 3.2 we introduce a novel technique to enhance the robustness of the calculations, which makes use of spatial information.

3.1. Minimum Statistics based robustness

The aim of this part is to detect frequency bins, which contain strong ratios of the near-end speech signal's power. Strong additive interferences make reliable estimations impossible and therefore, the computation of $\hat{\mathbf{D}}_i(m,l)$ will be halted at corrupted subbands containing measurable near-end speech signal power. As a first step, we estimate the magnitudes of the echo path transfer functions

$$|\hat{H}_i(m,l)|^2 \approx \frac{\hat{\Phi}_{Y_iY_i}(m,l)}{\hat{\Phi}_{XX}(m,l)} \approx \frac{\hat{\Phi}_{\Psi_i\Psi_i}(m,l) + \hat{\Phi}_{SS}(m,l)}{\hat{\Phi}_{XX}(m,l)}. \tag{12}$$

Since we presume that the echo path varies slowly, strong peaks in its estimate result from the near-end speech signal $S(m,l)$. We use minimum statistics to suppress these peaks (the operator 'MinStat' denotes the application of minimum statistics). Now, we can set up a condition to determine the presence of strong additive interferences

$$\frac{\hat{\Phi}_{\Psi_i\Psi_i}(m,l)}{\hat{\Phi}_{\Psi_i\Psi_i}(m,l)+\hat{\Phi}_{SS}(m,l)} \approx \frac{\text{MinStat}\left\{|\hat{H}_i(m,l)|^2\right\}}{|\hat{H}_i(m,l)|^2} < \mathcal{T}_{MS}. \quad (13)$$

$\mathcal{T}_{MS}$ is a threshold, which can be calculated by

$$\frac{1}{\left(\frac{\hat{\Phi}_{SS}(m,l)}{\hat{\Phi}_{\Psi\Psi}(m,l)}+1\right)} = \frac{1}{(\text{SER}(m,l)+1)} = \mathcal{T}_{MS}. \quad (14)$$

SER denotes the mean *speech-to-echo ratio*, which helps to find a suitable value for the threshold $\mathcal{T}_{MS}$. Frequency bins m_k, at which the condition in equation (13) is fulfilled, are not updated, since a reliable estimation of $\mathbf{D}_i(m,l)$ is not possible.

At low SERs, e.g. at 0 dB, we gain a solid robustness of the estimates against double-talk. However, the minimum statistics introduces a certain bias during the calculation of the nominator in equation (13). This could freeze the updating of $\hat{\mathbf{D}}_i(m,l)$ even at the absence of a near-end speech signal, when the SER is chosen too low. Hence, we have to face a trade-off between robustness against double-talk and fast estimation of the system misalignment transfer function.

3.2. Directivity Factor based robustness

Another possibility to enhance the robustness against double-talk represents the exploitation of spatial information. Therefore, we examine the so called *array gain* at the beamformer. With the help of the assumption in equation (6) it accounts to

$$G_A(m,l) = \frac{\text{SNR}_{\text{Array}}(m,l)}{\text{SNR}_{\text{Microphone}}(m,l)} \approx \frac{\bar{\Phi}_{\Xi\Xi}(m,l)}{\Phi_{\Xi_B\Xi_B}(m,l)}. \quad (15)$$

$\bar{\Phi}_{\Xi\Xi}(m,l)$ is the mean PSD gained by the residual echo in front of the beamformer. The mean can be calculated under the assumption of a homogeneous noise field generated by the residual echoes $\Xi_i(m,l)$. If we also suppose, that this noise field is diffuse, the array gain results into the so called *directivity factor* $\text{DF}(m)$ [6], which only depends on the beamformer's filter coefficients $A_i(m,l)$. Now, we can determine the residual echo's PSD by

$$\Phi_{\Xi_B\Xi_B}(m,l) = \text{DF}^{-1}(m)\bar{\Phi}_{\Xi\Xi}(m,l). \quad (16)$$

Figure 2 exemplarily shows the inverse of the directivity factor $\text{DF}^{-1}(m)$ in the dB-scale as a function of frequency. The sampling frequency f_s accounted to 8 kHz. We use a 4-microphone superdirective array in endfire steering with a spacing of 5 cm between adjacent microphones. The assumed signal-to-sensor noise ratio for a constraint of the array was set to 30 dB [6].

Let us now examine the ratio between the beamformer's input- and output-PSD

$$\frac{\Phi_{UU}(m,l)}{\bar{\Phi}_{EE}(m,l)} = \frac{\Phi_{\Xi_B\Xi_B}(m,l)+\Phi_{SS}(m,l)}{\bar{\Phi}_{\Xi\Xi}(m,l)+\Phi_{SS}(m,l)}. \quad (17)$$

If we introduce the *signal-to-residual echo ratio* (SRER)

$$\text{SRER}(m,l) = \frac{\Phi_{SS}(m,l)}{\bar{\Phi}_{\Xi\Xi}(m,l)}, \quad (18)$$

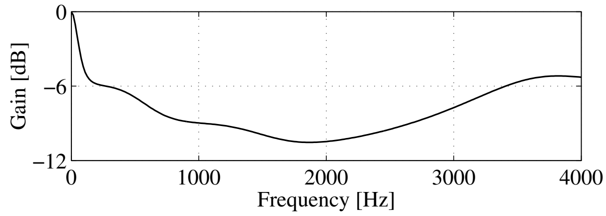

Fig. 2. Inverse of the directivity factor in the dB-scale as a function of frequency in Hz.

we can rewrite the ratio between the beamformer's input- and output-PSD and introduce a threshold $\mathcal{T}_{DF}$ in the same way as in section 3.1

$$\begin{aligned} & \frac{\text{DF}^{-1}(m)\bar{\Phi}_{\Xi\Xi}(m,l)+\Phi_{SS}(m,l)}{\bar{\Phi}_{\Xi\Xi}(m,l)+\Phi_{SS}(m,l)} \\ = \; & \frac{\text{DF}^{-1}(m)+\text{SRER}(m,l)}{1+\text{SRER}(m,l)} \\ > \; & \mathcal{T}_{DF}. \end{aligned} \quad (19)$$

At large SRERs, the quotient reaches values close to 1, oversteps the threshold $\mathcal{T}_{DF}$, and near-end speech activity is detected. At low SRERs, the quotient approaches $\text{DF}^{-1}(m)$. In figure 2 we can see that the directivity factor ends in 1 at small frequencies. Therefore, the newly proposed method will hardly work at very low frequencies. However, our exemplary array provides reliable results above 200 Hz.

4. SIMULATION RESULTS

In the following, we confirm our proposals by some simulation results. In the next section, we use white noise in order to compare the three investigated methods to estimate the residual echo according to section 2. Simulated RIRs at a length of 4096 with a reverberation time $\tau_{60} = 400$ ms come into operation. The RIRs are modified at sample 15,000. Directly after the microphones, there is one affine projection AEC for each microphone channel (projection order of 4, filter length of 512). Up from section 4.2, when double-talk is simulated as well, the AECs' adaptation is halted as soon as a near-end speaker starts to talk. The beamformer was designed as mentioned in section 3.2. The system misalignment estimation operates at a length of $L_{SME} = L'_{SME}L_{DFT} = 1024$.

4.1. Residual echo estimation methods

As already mentioned in section 2 the estimates using method 2 are biased. Both of the other methods deliver very similar results, which are biased at only 1 dB. All methods can follow the sudden modification of the RIR very quickly. Internal tests have shown that a single-channel estimation method delivers comparable results. For all further simulations we have chosen method 3, because it reveals good performance at "single-channel complexity".

4.2. Suppression of double-talk

In figure 4 we can see the impact of a near-end speech signal onto the estimation of the residual echo between sample 30,000 and

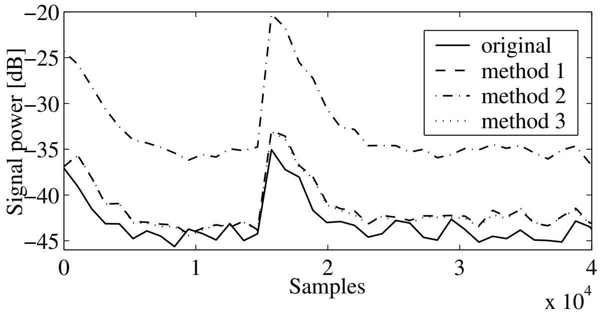

Fig. 3. Estimated residual echo signal powers and actual residual echo signal power ("original") as a function of time using white noise for the excitation signal $X(m, l)$.

50,000. Without any measures being taken the bias rises up to 25 dB. The SER to calculate the threshold $\mathcal{T}_{MS}$ for the minimum statistics based robustness was set to 6 dB. Still, there is a bias of about 15 dB. With an additional operation of the directivity factor based robustness (SRER of 0 dB to get $\mathcal{T}_{DF}$) the bias diminishes to 7 dB. Still, we can observe a quick and accurate reaction to the modification of the echo path at sample 15,000.

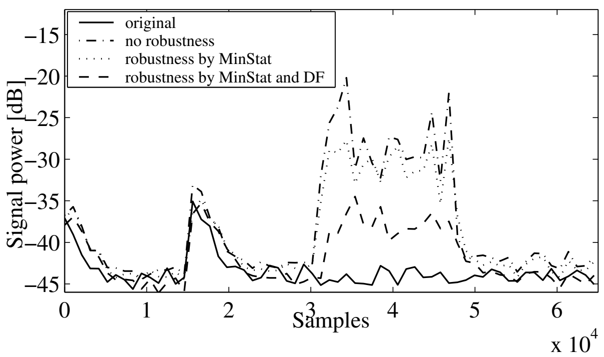

Fig. 4. Estimated residual echo signal powers and actual residual echo signal power as a function of time using white noise for the excitation signal $X(m, l)$ and a speech signal for $S(m, l)$.

4.3. Results with speech excitation

Instead of white noise we now use a real speech signal for the excitation signal $X(m, l)$. The near-end speech signal $S(m, l)$ is maintained. Between sample 30,000 and 40,000 there is a double-talk situation. Even the minimum statistics combined with the directivity factor based robustness cannot suppress all peaks, which are caused by the interferences. However, informal listening tests have shown that such over-estimations can hardly be heard, when a Wiener filter is applied at the beamformer's output (for audio samples, follow the www-link in [4]).

5. CONCLUSIONS

In this contribution we have proposed three methods to estimate the residual echo in a combined system with AECs running in parallel and a succeeding beamformer. Our simulation results show that the estimates are comparable to single-channel solutions as long as no near-end speaker is active. However, in double-talk periods the new multi-microphone approach increases robustness significantly. Informal listening test have revealed that there are no noticeable distortions of the near-end speech signal in such critical situations.

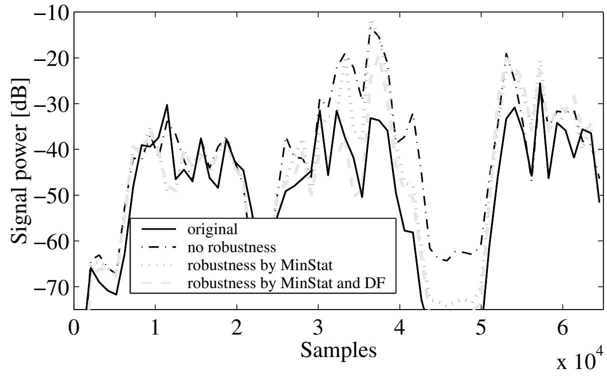

Fig. 5. Estimated residual echo signal powers and actual residual echo signal power as a function of time using a speech signal for the excitation signal $X(m, l)$.

6. REFERENCES

[1] W. L. Kellermann, "Acoustic Echo Cancellation for Beamforming Microphone Arrays," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. Ward, Eds., chapter 13, pp. 281–306. Springer-Verlag, 2001.

[2] 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.

[3] G. Enzner, R. Martin, and P. Vary, "Unbiased Residual Echo Power Estimation for Hands–Free Telephony," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, Orlando, Florida, USA, May 2002.

[4] M. Kallinger, J. Bitzer, and K. D. Kammeyer, "Residual Echo Estimation with the Help of Minimum Statistics," in *3rd IEEE Benelux Signal Processing Symposium*, Leuven, Belgium, Mar 2002, pp. 181–184, Can be downloaded via www.ant.uni-bremen.de/research/speech.

[5] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. Ward, Eds., chapter 3, pp. 39–60. Springer-Verlag, 2001.

[6] J. Bitzer and K. U. Simmer, "Superdirective microphone arrays," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. Ward, Eds., chapter 2, pp. 19–38. Springer-Verlag, 2001.

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