MULTICHANNEL ACOUSTIC ECHO SUPPRESSION

Karim Helwani[1], *Herbert Buchner*[2], *Jacob Benesty*[3], *and Jingdong Chen*[4]

[1]Quality and Usability Lab, Telekom Innovation Laboratories,[2] Machine Learning Group
[1,2]Technische Universität Berlin, 10587 Berlin, Germany
[3] INRS-EMT, University of Quebec, Montreal, QC H5A 1K6, Canada
[4] Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China

ABSTRACT

Acoustic echo suppression (AES) provides an attractive alternative to acoustic echo cancellation (AEC) techniques for full-duplex communication in low-complexity systems. However, so far AES techniques are commonly known to introduce significant distortions to the desired signal. Moreover, most traditional echo control techniques typically require accurately detecting the contribution of the near-end speaker to the microphone signal ("double talk"). The extension of AES techniques to the multichannel case usually assumes a symmetric system design which is often not fulfilled by typical scenarios. In this paper we propose a novel approach to multichannel acoustic echo suppression, which aims at extracting the near-end signal using a constraint for a distortionless output, without requiring a double-talk detector, or a symmetric system design. In addition to the above mentioned properties, the multichannel AES is also shown to overcome the known challenges in conventional multichannel acoustic echo control setups.

Index Terms— Acoustic echo suppression, multichannel adaptive filtering, minimum variance distortionless response filter.

1. INTRODUCTION

Multichannel sound reproduction enhances realism in virtual reality and multimedia communication systems. In hands-free multichannel communication setups, disturbing echoes are produced by the acoustic feedback of the loudspeakers' signals into the microphones. AEC aims at canceling the acoustic echoes from the microphone signals. In a typical multichannel AEC with P reproduction channels and a single microphone channel in the receiving (near-end) room, the signals of the P reproduction channels originate from speech- or audio sources at the far-end.

To cancel the echoes arising due to the acoustic path in the near-end, the reproduction signals $x_p(t)$ are filtered with the adaptively estimated $P \cdot L$ coefficients of the FIR filter $\hat{\mathbf{g}} = [\hat{\mathbf{g}}_1^{\mathrm{T}}, \ldots, \hat{\mathbf{g}}_P^{\mathrm{T}}]^{\mathrm{T}}$, i.e., a replica of the actual acoustic multiple-input single-output (MISO) system. The resulting signal $\hat{y}(t)$ is subtracted from the near-end microphone signal $d(t)$, where t denotes the time instant. If the estimated echo paths $\hat{\mathbf{g}}$ are equal to the true transfer paths $\mathbf{g}$, all disturbing echoes will be canceled from the microphone signal. Note, that the multiple-input multiple-output (MIMO) case can be considered as multiple parallel independent MISO systems for each microphone channel. Hence, the consideration of a MISO system in the near-end room is sufficient in the context of this work.

In acoustic echo control, residual echo suppressors, originally introduced in a heuristic way, are typically employed after the actual system identification-based AEC in order to meet the requirements for a high attenuation of the echoes in practical applications including, e.g., quickly time-varying acoustic environments, microphone noise, and considerable network delay [1, 2]. As an extreme case, under the assumption of a simplified echo path model consisting of delay and short-time spectral modification, a system purely based on the residual echo suppression stage (acoustic echo suppression, AES) has been proposed in [3, 4, 5, 6, 7, 8].

The basic notion of AES is a spectral modification of the microphone signal $d(t)$ in order to attenuate its echo component that is caused by the acoustical feedback of the loudspeaker signal $x(t)$ along the unknown echo path. The core assumption which has been made in [6], is that the echo path (room impulse response) can entirely be modeled by a linear phase filter, i.e., on its way to the microphone, the loudspeaker signal is shifted in time and its magnitude spectrum is shaped. The latter effect, also called coloration, is mostly caused by early reflections of the room. Hence, in this model the impact of late reflections is ignored.

Once the delay has been estimated, a coloration filter can be derived based on the Wiener filtering approach. The suppression filter is then designed to be orthogonal to the signal representing the divergence of the estimated signal using the coloration filter and the amplitude of the near-end signal. AEC algorithms for the multichannel case often suffer from the fact that the signals of the multichannel reproduction system are usually not only intrachannel correlated but typically also highly interchannel correlated. This results in an ill-conditioned correlation matrix in the underlying normal equation of the MISO adaptive filter. Strategies to cope with the mentioned ill-conditioning problem aim either at enhancing the conditioning by manipulating the input signals, as long as the manipulation can be perceptually tolerated [9, 10], or at regularizing the problem to determine an approximate solution that is stable under small changes in the initial data [11, 12, 13].

The extension of the AES approach to the multichannel case in [8] is based on summing up the loudspeaker signals into one signal $\sum_{p=1}^{P} x_p(t)$ and then treating the MISO case as the SISO case. This simplification inherently assumes a symmetric system setup such that all loudspeaker signals have the same delay at the microphone. Moreover, suppression techniques are commonly known to introduce distortions to the desired signal. Moreover, AEC as well as the briefly reviewed AES typically require accurately detecting the contribution of the near-end speaker to the microphone signal ("double talk").

This paper addresses both the distortion and double-talk problems. In order to limit the signal distortion to a minimum in AES systems, we present in this paper a novel two-stage approach which explicitly constrains the near-end signal. Using the interframe statistics of the signal and extending the work in [14, 15] allow us to derive a suit-

ably designed minimum variance distortionless response (MVDR) filter. Similar to our previous work [16], the presented echo control system does not require double-talk detection.

2. PROBLEM FORMULATION AND THE PROPOSED APPROACH

2.1. Signal Model

Let us consider the conventional signal model in which acoustic echoes are generated from the coupling between P loudspeakers and a microphone. The microphone signal at the time index t can be written as

$$\begin{aligned} d(t) &= \sum_{p=1}^{P} g_p(t) * x_p(t) + u(t) \\ &= y(t) + u(t), \end{aligned} \tag{1}$$

where $x_p(t)$ is the p-th loudspeaker (or far-end) signal, $g_p(t)$ is the impulse response from the p-th loudspeaker to the microphone, $u(t)$ is the near-end signal, and $y(t)$ is the echo signal. We assume that $y(t)$ and $u(t)$ are uncorrelated. All signals are considered to be real, zero mean, and broadband. Using the short-time Fourier transform (STFT), Eq. (1) can be expressed in the time-frequency domain as

$$D(k,n) = Y(k,n) + U(k,n), \tag{2}$$

where $D(k,n)$, $Y(k,n)$, and $U(k,n)$ are the STFTs of $d(t)$, $y(t)$, and $u(t)$, respectively, at the frequency bin $k \in \{0,1,\ldots,K-1\}$ and the time frame n. Later on, the approximation of the echo signal:

$$\begin{aligned} Y(k,n) &\approx \begin{bmatrix} G_1^*(k) & G_2^*(k) & \ldots & G_P^*(k) \end{bmatrix} \cdot \begin{bmatrix} X_1(k,n) \\ X_2(k,n) \\ \vdots \\ X_P(k,n) \end{bmatrix}, \\ &= \mathbf{G}^{\mathrm{H}}(k,n) \cdot \mathbf{X}(k,n), \end{aligned} \tag{3}$$

will be used, where $G(k)$ and $X(k,n)$ are the STFTs of $g(t)$ and $x(t)$, and superscript $\{\cdot\}^*$ is the complex-conjugate operator. Hence, the microphone signal can be described as

$$D(k,n) = \begin{bmatrix} \mathbf{G}^{\mathrm{H}}(k) & 1 \end{bmatrix} \begin{bmatrix} \mathbf{X}(k,n) \\ U(k,n) \end{bmatrix}. \tag{4}$$

Further, we assume that the near-end and echo signal are uncorrelated such that

$$\hat{\mathcal{E}}\{U(k,n)X_p^*(k,n)\} = 0 \qquad \forall p \in \{1,\ldots,P\}, \tag{5}$$

where $\hat{\mathcal{E}}\{\cdot\}$ denotes an empirical value of the expectation.
In the following section, we introduce a solution based on the shown assumptions (4) and (5), and composed of two processing stages as depicted in Fig. 1. In the first stage, an initial guess of the near-end signal is obtained. The estimated signal is then post-processed in terms of minimizing the distortions.

2.2. Initial Guess of the Near-End Signal

For simultaneous estimation of $\mathbf{G}(k)$, and the near-end signal $U(k,n)$, we set up the following system of equations by combining Eq. (4) and (5):

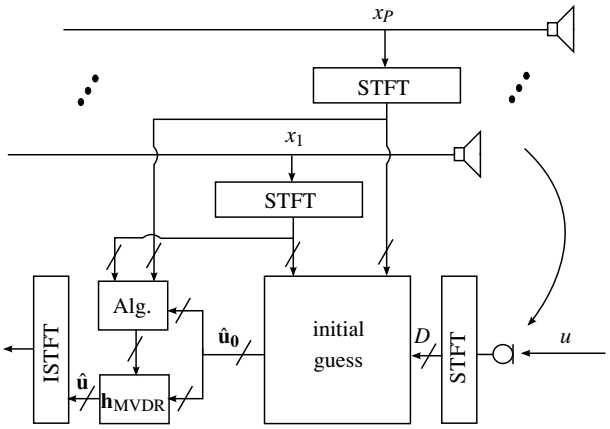

Fig. 1. Block diagram of the proposed system.

$$\begin{aligned} &\begin{bmatrix} \mathbf{d}(k,n) \\ \hline \mathbf{0}_{M_1\times 1} \end{bmatrix} = \\ &\left[\begin{array}{c|cc} \mathcal{X}'(k,n) & \multicolumn{2}{c}{\mathbf{I}_{M_2\times M_2}} \\ \hline \mathbf{0}_{M_1\times P} & \mathrm{circ}(\mathcal{X}^{\mathrm{H}})(k,n) & \mathbf{0}_{M_1\times(M_2-M_1)} \end{array}\right] \cdot \begin{bmatrix} \hat{\mathbf{G}}^*(k) \\ \hline \hat{\mathbf{u}}_{\mathbf{0}}(k,n) \end{bmatrix}, \end{aligned} \tag{6}$$

where
$$\begin{aligned} \mathcal{X}'(k,n) &:= [\mathbf{X}(k,n),\ldots,\mathbf{X}(k,n-M_2+1)]^{\mathrm{T}}, \\ \mathbf{d}(k,n) &:= [D(k,n),D(k,n-1),\ldots,D(k,n-M_2+1)]^{\mathrm{T}}, \\ \mathcal{X}(k,n) &:= [\mathbf{X}(k,n),\ldots,\mathbf{X}(k,n-M_1+1)]^{\mathrm{T}}, \end{aligned}$$

$$\mathrm{circ}(\mathcal{X}^{\mathrm{H}})(k,n) := \begin{bmatrix} \mathbf{X}^*(k,n) & \mathbf{X}^*(k,n-1) & \ldots & \mathbf{X}^*(k,n-M_1+1) \\ \mathbf{X}^*(k,n-M_1+1) & \mathbf{X}^*(k,n) & \ldots & \mathbf{X}^*(k,n-M_1+2) \\ \vdots & \ddots & & \vdots \\ \mathbf{X}^*(k,n-1) & \mathbf{X}^*(k,n-2) & \ldots & \mathbf{X}^*(k,n) \end{bmatrix},$$

$$\hat{\mathbf{u}}_{\mathbf{0}}(k,n) := [\hat{U}_0(k,n),\ldots,\hat{U}_0(k,n-M_2+1)]^{\mathrm{T}},$$

which is an estimate of

$$\mathbf{u}(k,n) := [U(k,n),\ldots,U(k,n-M_2+1)]^{\mathrm{T}}.$$

$\hat{\mathbf{u}}_{\mathbf{0}}$ can be obtained from Eq. (6) by the pseudoinverse.
Note that the matrix on the right-hand side in (6) exclusively depends on the loudspeaker signals $\mathbf{X}(\cdot)$, while the left-hand side exclusively depends on the microphone signal $D(\cdot)$. The solution of Eq. (6) can be interpreted as an explicit block-online version of [16], explaining that this approach works without additional double-talk detection.

2.3. Complexity Reduction for the Massive Multichannel Case

In multichannel reproduction techniques, such as Stereo, 5.1 surround sound, and wave field synthesis (WFS) techniques, the loudspeakers emit highly crosscorrelated signals, e.g., the impulse responses of a WFS system rendering one point source are nearly

unit impulses with different, suitably chosen delays and amplitudes. Therefore, the P-dimensional vector $\mathbf{X}(k,n)$ representing the loudspeaker signals can be transformed into a lower dimensional $\tilde{\mathbf{X}}(k,n)$ using a transformation matrix $\mathbf{T}(k,n)$ containing the orthogonal vectors spanning the eigenspace of the signal [17]. These can be obtained as the eigenvectors of the following matrix

$$\mathbf{R}_{\mathbf{xx}}(k,n) := \quad \mathbf{R}_{\mathbf{xx}}(k,n-1) + \mathbf{X}(k,n)\mathbf{X}^{\mathrm{H}}(k,n), \tag{7}$$

where is a forgetting factor. The square $P \times P$ matrix $\mathbf{R}_{\mathbf{xx}}(k,n)$ can be decomposed into

$$\mathbf{R}_{\mathbf{xx}}(k,n) = \mathbf{T}'(k,n)\tilde{\mathbf{R}}_{\mathbf{xx}}(k,n)\mathbf{T}'^{\mathrm{H}}(k,n), \tag{8}$$

with $\mathbf{T}'(k,n)\mathbf{T}'^{\mathrm{H}}(k,n) = \mathbf{I}$ where $\mathbf{I}$ is the unity matrix, and $\tilde{\mathbf{R}}_{\mathbf{xx}}(k,n)$ is a diagonal matrix.
Let us define $\mathbf{T}(k,n)$ as the submatrix with the dimensions $P \times R$ containing the R eigenvectors corresponding to the largest $R \leq P$ eigenvalues. Note, that due to the iterative estimation of the autocorrelation matrix, its eigenvalue decomposition can be efficiently computed [18, 19].
Further, We define

$$\tilde{\mathbf{X}}(k,n) := \mathbf{T}^{\mathrm{H}}(k,n)\mathbf{X}(k,n), \quad \tilde{\hat{\mathbf{G}}}(k,n) := \mathbf{T}^{\mathrm{H}}(k,n)\hat{\mathbf{G}}(k,n). \tag{9}$$

Since the vector $\mathbf{X}$ is optimally embedded in the space spanned by the column vectors of $\mathbf{T}$ it can easily be verified that

$$Y(k,n) \approx \tilde{\hat{G}}^{\mathrm{H}}(k,n) \cdot \tilde{\mathbf{X}}(k,n). \tag{10}$$

Hence, the use of the transformed quantities allow us to set up a system of equations for simultaneous estimation of $\mathbf{G}(k)$, and the near-end signal $U(k,n)$, which is typically much smaller than Eq. (6). In a typical full-duplex communication setup using a WFS system P could lie up to several hundreds and R depends on the active sources in the far-end, e.g., one or two speakers. In (6), we make the replacements $\mathcal{X}'(k,n) \rightarrow \tilde{\mathcal{X}}'(k,n)$, $\mathcal{X}(k,n) \rightarrow \tilde{\mathcal{X}}(k,n)$, where $\tilde{\mathcal{X}}$ and $\tilde{\mathcal{X}}'$ are built up analogously to $\mathcal{X}$ and $\mathcal{X}'$ but using the transformed loudspeaker signals as given in (9). Further, we replace $\mathbf{0}_{M_1 \times P}$ by $\mathbf{0}_{M_1 \times R}$, and $\hat{\mathbf{G}}^*(k)$ by $\tilde{\hat{\mathbf{G}}}^*(k)$.

3. MVDR PROCESSING STAGE

The elements $\hat{U}_0(k,n)$, could still contain both a residual echo component that is considered as an interference and a part of the desired near-end signal.
For a suppression of the residual echo signal we consider further decomposing the estimated near-end signal as follows:

$$\hat{\mathbf{u}}_{\mathbf{0}}(k,n) = \mathbf{u}_{\mathrm{c}}(k,n) + \mathbf{u}_{\mathrm{i}}(k,n) + \mathbf{r}(k,n), \tag{11}$$

where $\mathbf{r}$ denotes the residual echo, $\mathbf{u}_{\mathrm{c}}$ is the component of the estimated near-end signal vector which is coherent with $U(k,n)$, and $\mathbf{u}_{\mathrm{i}}$ is the incoherent component, that is orthogonal to the coherent component $\mathbf{u}_{\mathrm{c}}$. In the following we show how the decomposition in Eq. (11) can be done in practice by deriving a MVDR filter for the estimated near-end signal. The idea is to estimate a distortionless version $\hat{U}(k,n)$ of the near-end signal starting from the initial estimation $\hat{\mathbf{u}}_{\mathbf{0}}(k,n)$. Coherence between $U(k,n)$ and $\hat{U}(k,n)$ occurs if the following condition is fulfilled

$$\hat{\mathcal{E}}\{\hat{U}(k,n)U^*(k,n)\} \stackrel{!}{=} \ _{U}(k,n), \tag{12}$$

where

$$_{U}(k,n) := \hat{\mathcal{E}}\{U(k,n)U^*(k,n)\}. \tag{13}$$

Using

$$\begin{aligned}\hat{U}(k,n) =& \mathbf{h}^{\mathrm{H}}(k,n)\hat{\mathbf{u}}_{\mathbf{0}}(k,n)\\ =& \mathbf{h}^{\mathrm{H}}(k,n)[\mathbf{u}_{\mathrm{c}}(k,n) + \mathbf{u}_{\mathrm{i}}(k,n) + \mathbf{r}(k,n)],\end{aligned} \tag{14}$$

we obtain with $\mathbf{u}_{\mathrm{c}}(k,n) = \ _{\mathbf{u}}(k,n) \cdot U(k,n)$ and (12)

$$\begin{aligned}\hat{\mathcal{E}}\{\hat{U}(k,n)U^*(k,n)\} =& \mathbf{h}^{\mathrm{H}}(k,n)\hat{\mathcal{E}}\{\mathbf{u}_c(k,n)U^*(k,n)\}\\ =& \mathbf{h}^{\mathrm{H}}(k,n) \ _{\mathbf{u}}(k,n)\hat{\mathcal{E}}\{\hat{U}(k,n)U^*(k,n)\}.\end{aligned} \tag{15}$$

For determining $_{\mathbf{u}}(k,n)$ we derive

$$\begin{aligned}\hat{\mathcal{E}}\{\mathbf{u}_c(k,n)U^*(k,n)\} =& \hat{\mathcal{E}}\{\mathbf{u}(k,n)U^*(k,n)\}\\ =& \ _{\mathbf{u}}(k,n)\hat{\mathcal{E}}\{U(k,n)U^*(k,n)\},\end{aligned} \tag{16}$$

$$_{\mathbf{u}}(k,n) = \frac{\hat{\mathcal{E}}\{\mathbf{u}(k,n)U^*(k,n)\}}{_{U}(k,n)}. \tag{17}$$

Note, that $_{\mathbf{u}}(k,n)$ can be understood as a weighted version of the single eigenvector of the rank-one matrix $\mathbf{u}\mathbf{u}^{\mathrm{H}}$. Now, from condition (15) we immediately obtain the following important constraint for $\mathbf{h}$ to estimate the near-end signal with no distortion:

$$\mathbf{h}^{\mathrm{H}}(k,n) \ _{\mathbf{u}}(k,n) = 1. \tag{18}$$

In the practical implementation we determine $_{\mathbf{u}}(k,n)$ using the initial guess $\hat{\mathbf{u}}_{\mathbf{0}}$. In Eq. (11), $\mathbf{r}$ in turn can be decomposed into two distinct parts: a coherent one and an incoherent one relative to the echo signal. In general, a constraint can be added to minimize the residual echo by choosing $\mathbf{h}$ to be additionally orthogonal to the subspace spanned by the loudspeaker signals. But here, the solution of the system of equations in Eq. (6) offers in practice an almost echo free estimation of the near-end signal such that applying further constraints does not yield in statistically significant improvement of the attenuation of the echo.

3.1. Minimum Variance

Based on the minimum variance criterion, we aim at minimizing the cost function:

$$\begin{aligned}J_0(\mathbf{h}) :=& \hat{\mathcal{E}}\{\hat{U}(k,n)\hat{U}^*(k,n)\}\\ =& \mathbf{h}^{\mathrm{H}}\hat{\mathcal{E}}\{\hat{\mathbf{u}}_{\mathbf{0}}(k,n)\hat{\mathbf{u}}_{\mathbf{0}}^{\mathrm{H}}(k,n)\}\mathbf{h} = \mathbf{h}^{\mathrm{H}} \ _{\hat{\mathbf{u}}_{\mathbf{0}}\hat{\mathbf{u}}_{\mathbf{0}}}\mathbf{h}.\end{aligned} \tag{19}$$

By assuming a prior multivariate normal distribution with zero mean for $\mathbf{h}$ we obtain one more constraint on the ℓ_2-norm of $\mathbf{h}$. The regularized cost function reads

$$J_1(\mathbf{h}) := \mathbf{h}^{\mathrm{H}} \ _{\hat{\mathbf{u}}_{\mathbf{0}}\hat{\mathbf{u}}_{\mathbf{0}}}\mathbf{h} + \ \mathbf{h}^{\mathrm{H}}\mathbf{h}. \tag{20}$$

3.2. Distortionless Response

The constraint in Eq. (18) can be added to the cost function Eq. (19) using the Lagrangian multiplier technique yielding the new cost function:

$$J(\mathbf{h}) := \mathbf{h}^{\mathrm{H}} \ _{\hat{\mathbf{u}}_{\mathbf{0}}\hat{\mathbf{u}}_{\mathbf{0}}}\mathbf{h} + \ \mathbf{h}^{\mathrm{H}}\mathbf{h} + \ (1 - \ _{\mathbf{u}}^{\mathrm{H}}\mathbf{h}). \tag{21}$$

At the minimum the gradient of the cost function is zero and we derive after several straightforward calculation steps:

$$\mathbf{h}_{\mathrm{MVDR}}(k,n) = (\quad \mathbf{u_0 u_0} + \quad \mathbf{I})^{-1} \quad \mathbf{u} \left[\quad \overset{\mathrm{H}}{\mathbf{u}} (\quad \hat{\mathbf{u}}_0 \hat{\mathbf{u}}_0 + \quad \mathbf{I})^{-1} \quad \mathbf{u} \right]^{-1}. \tag{22}$$

4. EXPERIMENTAL RESULTS

4.1. Performance Measures

The two most important means to evaluate the acoustic echo suppression performance are the attenuation of the acoustic echo, and the distortion of the near-end signal. We define the fullband acoustic echo reduction factor at the time frame n as

$$(n) = \frac{\sum_{k=0}^{K-1} \quad _Y(k,n)}{\sum_{k=0}^{K-1} \quad _{\hat{U}}(k,n)}, \tag{23}$$

where $\quad _Y(k,n)$, and $\quad _{\hat{U}}(k,n)$ are defined analogously to Eq. (13). The acoustic echo reduction factor should be greater than or equal to 1. When $\quad = 1$, there is no echo reduction and the higher the value of $\quad$, the more the echo is reduced. This definition is equivalent to the echo-return loss enhancement (ERLE) [20]. Further, we define the fullband near-end signal distortion index at the time frame n as

$$v(n) := \frac{\sum_{k=0}^{K-1} \hat{\mathcal{E}}\{|\hat{U}(k,n) - U(k,n)|^2\}}{\sum_{k=0}^{K-1} \quad _U(k,n)}. \tag{24}$$

4.2. Simulations

To evaluate how successful the described algorithm is in suppressing the echo signal, three experiments were conducted. In the first simulation only a (female) far-end speaker is talking. The signal is reproduced in the near-end room using 2, 5, and 7 loudspeakers respectively. The far-end room is simulated using measured impulse responses of a room with a reverberation time (T_{60}) of approximately 200 ms. The measured impulse response of the near-end room exhibit $T_{60} \approx 400$ ms. In each loudspeaker setup the loudspeaker signals are normalized such that the RMS of the microphone signal is independent from the loudspeaker number. To make the setting more realistic, Gaussian white noise is added to the microphone signal with an SNR of 35 dB relative to the RMS of the signal at the microphone. The sampling frequency of the signals is 8 kHz. The chosen DFT length is 256 with an overlap factor of 50%. The filter length was set to $M_1 = M_2 = 8$.

The position of the rendered virtual source was changed one time at $t \approx 3.9$ s by changing the set of the impulse responses of the far-end (the accurate instant is marked by the vertical line). The achieved echo return loss enhancement is shown in Fig. 2. Simulations show that the echo suppression is nearly independent of the channel number. Moreover, changing the impulse responses in the far-end does not lead to breaking down the achieved ERLE as it is the case in typical AEC algorithms without applying preprocessing techniques [9].

In the second experiment both speakers talk simultaneously ("double talk"). Far-end and near-end speech signals have been adjusted manually to exhibit roughly equal loudness, the distortion of the extracted near-end signal is shown in Fig. 3 for different filter lengths $M_1 = M_2 \in \{2, 4, 8, 16\}$. The distortion of the near-end signal in the double-talk period is upper limited to -15 dB and is as expected, even better in the case of only the (male) speaker at the near-end is active, as the results given in Fig. 4 show.

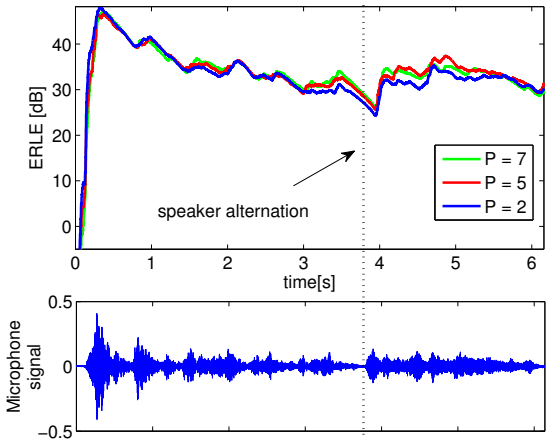

Fig. 2. Achieved echo-return loss enhancement of the proposed system in the single-talk period for different numbers of channels.

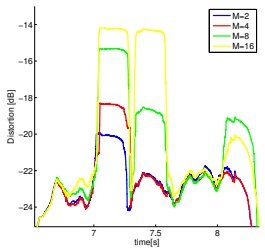

Fig. 3. Achieved distortion of the near-end signal during the double-talk period.

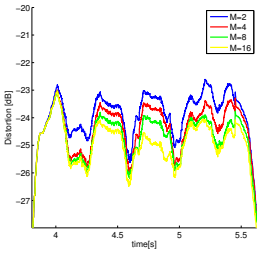

Fig. 4. Achieved distortion of the near-end signal during the period where only the near-end speaker is active.

5. CONCLUSION

In this paper, we presented an approach to multichannel acoustic echo suppression, which extracts the near-end signal from the microphone signal with a distortionless constraint and without requiring a double-talk detector. The new approach offers high degrees of flexibility, is scalable and highly efficient as the presented simulation results have shown.

6. RELATION TO PRIOR WORK

The single-channel formulation for AES presented in [3, 7] has been extended to the multichannel case in [4, 8]. The approach in [4] requires decorrelating the loudspeaker signal by a preprocessing stage like traditional multichannel AEC. The approach in [8] requires inherently a symmetric system design and an accurate delay estimation. Both approaches require a double-talk detector and are known to introduce distortion to the desired near-end signal. The presented approach in this paper copes with highly correlated loudspeaker signals of multichannel reproduction systems, does not require a double-talk detector, and constrains near-end signal distortion.

7. REFERENCES

[1] R. Martin and J. Altenhoner, "Coupled adaptive filters for acoustic echo control and noise reduction," in *Proc. IEEE ICASSP*, 1995, vol. 5, pp. 3043–3043.

[2] G. Enzner, H. Buchner, A. Favrot, and F. Kuech, "Acoustic echo control," in *R. Chellappa and S. Theodoridis (eds.), Electronic Reference in Signal, Image, and Video Processing*. Elsevier/Academic Press, 2013.

[3] C. Avendano, "Acoustic echo suppression in the STFT domain," in *Proc. IEEE WASPAA*, 2001, pp. 175–178.

[4] C. Avendano and G. Garcia, "STFT-based multi-channel acoustic interference suppressor," in *Proc. IEEE ICASSP*, 2001, vol. 1, pp. 625–628.

[5] C. Faller and J. Chen, "Suppressing acoustic echo in a spectral envelope space," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 5, pp. 1048– 1062, Sept. 2005.

[6] C. Faller and C. Tournery, "Estimating the delay and coloration effect of the acoustic echo path for low complexity echo suppression," in *Proc. IWAENC*, 2005, pp. 1–4.

[7] C. Faller and C. Tournery, "Robust acoustic echo control using a simple echo path model," in *Proc. IEEE ICASSP*, 2006, vol. 5, pp. 281–284.

[8] C. Faller and C. Tournery, "Stereo acoustic echo control using a simplified echo path model," in *Proc. IWAENC*, 2006, pp. 1–4.

[9] J. Benesty, D.R. Morgan, and M.M. Sondhi, "A better understanding and an improved solution to the specific problems of stereophonic acoustic echo cancellation," *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 2, pp. 156–165, 1998.

[10] J. Herre, H. Buchner, and W. Kellermann, "Acoustic echo cancellation for surround sound using perceptually motivated convergence enhancement," in *Proc. IEEE ICASSP*, 2007, vol. 1, pp. 17–20.

[11] H. Buchner, S. Spors, and W. Kellermann, "Wave-domain adaptive filtering: acoustic echo cancellation for full-duplex systems based on wave-field synthesis," in *Proc. IEEE ICASSP*, 2004, vol. 4, pp. 117–120.

[12] K. Helwani, H. Buchner, and S. Spors, "Source-domain adaptive filtering for MIMO systems with application to acoustic echo cancellation," in *IEEE ICASSP*, 2010, pp. 321–324.

[13] K. Helwani, H. Buchner, and S. Spors, "Multichannel adaptive filtering with sparseness constraints," in *Proc. IWAENC*, 2012, pp. 1–4.

[14] J. Benesty and Y. Huang, "A single-channel noise reduction MVDR filter," in *Proc. IEEE ICASSP*, 2011, pp. 273 –276.

[15] J. Benesty, J. Chen, and E.A.P. Habets, *Speech Enhancement in the STFT Domain*, Berlin, Germany: Springer-Verlag, 2011.

[16] H. Buchner and W. Kellermann, "A fundamental relation between blind and supervised adaptive filtering illustrated for blind source separation and acoustic echo cancellation," in *Proc. HSCMA*, 2008, pp. 17 –20.

[17] S. Spors, H. Buchner, and K. Helwani, "Block-based multichannel transform-domain adaptive filtering," in *Proc. EUSIPCO*, 2009, pp. 1735–1739.

[18] J.R. Bunch, Ch.P. Nielsen, and D.C. Sorensen, "Rank-one modification of the symmetric eigenproblem," *Numerische Mathematik*, vol. 31, no. 1, pp. 31–48, 1978.

[19] K. Helwani, H. Buchner, and S. Spors, "On the robust and efficient computation of the kalman gain for multichannel adaptive filtering with application to acoustic echo cancellation," in *Proc. 44-th Asilomar Conference on Signals, Systems and Computers*, 2010, pp. 988–992.

[20] J. Benesty, T. Gänsler, D.R. Morgan, M.M. Sondhi, and S.L. Gay, *Advances in Network and Acoustic Echo Cancellation*, Springer-Verlag Berlin Heidelberg, 2001.