FULL-DUPLEX COMMUNICATION SYSTEMS USING LOUDSPEAKER ARRAYS AND MICROPHONE ARRAYS

Herbert Buchner, Sascha Spors, Walter Kellermann, and Rudolf Rabenstein

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

ABSTRACT

For high-quality multimedia communication systems, such as teleconferencing, or tele-teaching (especially of music), multichannel sound reproduction is highly desirable. While current approaches still rely on a restrained listening area, the sweet spot, a volume solution for a large listening space is offered by the Wave Field Synthesis (WFS) method, where arrays of loudspeakers generate a prespecified soundfield. On the recording side of the two-way systems, the use of microphone arrays is an effective approach to cope with undesired signal components in the receiving room. However, before full-duplex communication can be deployed, efficient approaches to the acoustic echo cancellation (AEC) problem in this challenging scenario have to be found. In this paper, we investigate different options for system integration, after a brief discussion of the current state of the art. We then present a first real-time solution on a regular PC platform, based on an efficient AEC for MIMO (multi-input and multi-output) systems in the frequency-domain.

1. INTRODUCTION

To enhance the sound realism in multimedia communication systems, such as teleconferencing or tele-teaching (especially of music), and to create a three-dimensional illusion of sound sources positioned in a virtual acoustical environment, multichannel sound reproduction is necessary. However, advanced loudspeaker-based approaches, like the 3/2-Surround format still rely on a restrained listening area, the sweet spot. A volume solution for a large listening space is offered by the Wave Field Synthesis (WFS) method [7], where arrays of a large number of independently driven loudspeakers generate a prespecified soundfield.

On the recording side of the two-way systems, the use of microphone arrays is an effective approach to separate desired and undesired sources in the receiving room, and to cope with reverberation of the recorded signal. *Figure* 1 shows the general setup for such a system.

However, before full-duplex communication can be deployed, the problem of acoustic feedback from the P loudspeakers to the Q microphones has to be addressed. Acoustic echo cancellation (AEC) for the resulting $P \cdot Q$ echo paths (where P usually lies between 20 and several hundred, and Q may be on the order of 10 to 100) poses the major problem in this MIMO (multi-input and multi-output) context, and satisfactory solutions for AEC in conjunction with large loudspeaker arrays have not been presented.

This work was partly supported by grants from Grundig AG, Nuremberg leading the German EMBASSI consortium, and from the European Commission as sponsor of the CARROUSO project.

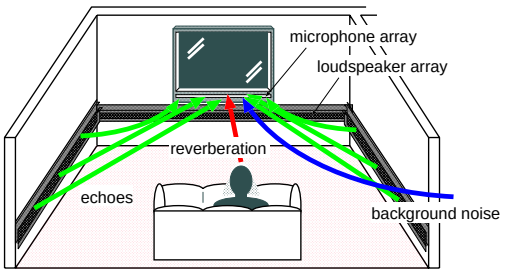

Fig. 1. General setup for multimedia communication.

In this paper, we first review the state-of-the art of multichannel acoustic echo cancellation. Then, from the single components of the communication system, we establish a generic matrix framework. This framework will be used to discuss different strategies and limitations for system integration, followed by an application of a recently introduced and efficient MIMO-system acoustic echo cancellation. As we will see, the generic framework gives way to several suboptimum strategies, while the generality of the proposed MIMO AEC allows flexibility for many potential applications. A first real-time solution has been implemented entirely on a regular PC platform.

2. AEC - STATE OF THE ART

Classical AEC applications are hands-free telephony or teleconference systems, where most of them are still based on monaural sound reproduction. Only recently, first stereophonic prototypes appeared, and lately, it has become possible to realize such systems on regular PCs using an efficient frequency-domain framework [1], [2]. Moreover, we efficiently extended the system to the multichannel case, and could successfully demonstrate its applicability for 5-channel surround sound [3], [4].

The fundamental idea of any P-channel AEC structure (*Fig.* 2) is to use adaptive FIR filters with impulse response vectors $\hat{\boldsymbol{h}}_p(k)$, $p = 1, \ldots, P$ that identify the truncated (generally time-varying) echo path impulse responses $\boldsymbol{h}_p(k)$. The filters $\hat{\boldsymbol{h}}_p(k)$ are stimulated by the loudspeaker signals $x_p(k)$ and, then, the resulting echo estimates $\hat{y}_p(k)$ are subtracted from the microphone signal $y(k)$ to cancel the echoes.

The specific problems of MC AEC include all those known

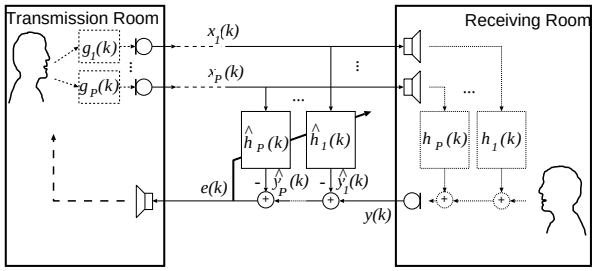

Fig. 2. Basic MC AEC structure

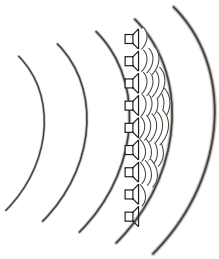

Fig. 3. Basic principle of Wave Field Synthesis

for mono AEC (e.g. [5]), but in addition to that, MC AEC often 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 [6]. The correlation results from the fact that the signals are almost always derived from common sound sources in the transmission room, as shown in *Fig.* 2. Straightforward extension of known mono AEC schemes thus often leads to very slow convergence of the adaptive filter towards the physically true echo paths [5, 6]. If the relation between the signals $x_p(k)$ is strictly linear, then there is a fundamental problem of non-uniqueness in the multichannel case as was shown in [6]. 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 [6]. The problem can be relieved by some nearly 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 MC AEC [5].

Before introducing the full MIMO AEC scheme in section 6, we first discuss the loudspeaker reproduction system in section 3, the microphone recording system in section 4, and show strategies for system integration in section 5.

3. WAVE FIELD SYNTHESIS

Wave field synthesis (WFS) using loudspeaker arrays is based on Huygens principle. It states that any point of a wave front of a propagating wave at any instant conforms to the envelope of spherical wavelets emanating from every point on the wavefront at the prior instant. This principle can be used to synthesize acoustical wavefronts of arbitrary shape. The mathematical foundation is given by the Kirchhoff-Helmholtz integrals [7]. They state that at any listening point within the source-free volume the sound pressure can be calculated if both the sound pressure and its gradient are known on the surface enclosing the volume. For practical implementations the surface degenerates to a line surrounding the listening area and the acoustic sources are realized by loudspeakers on discrete positions on this line. Figure 3 illustrates this principle. When WFS is used for sound reproduction in the receiving room, then the basic MC AEC structure from *Fig.* 2 does not apply any more. Instead of recording and transmitting P loudspeaker signals as in *Fig.* 2, only a small number of $P'' << P$ source signals is transmitted. From these, the loudspeaker signals are generated by a two step procedure: aurealization of the transmission room or an arbitrary virtual room and compensation of the acoustics in the receiving room.

Aurealization is performed by convolution of the source signals $x''(k)$ with a matrix $\boldsymbol{A}(k)$ of impulse responses

$$\boldsymbol{x}'(k) = \boldsymbol{A}(k) * \boldsymbol{x}''(k). \tag{1}$$

The room impulse responses in $\boldsymbol{A}$ are either recorded in advance by special microphone arrays as shown in [8] or computed according to the WFS theory [7].

Having now created virtual room acoustics by matrix $\boldsymbol{A}$, we also need to take into account the actual room impulse responses $\boldsymbol{H}_L$ in the receiving room from the array to the listener. This is done by a bank of suitable prefilters $\boldsymbol{G}$ that are designed so that, ideally, the influence of $\boldsymbol{H}_L$ is equalized [9]

$$\|\boldsymbol{H}_L(k) * \boldsymbol{G}(k) - \boldsymbol{F}(k)\| \rightarrow \min, \tag{2}$$

where $\boldsymbol{F}(k)$ is an appropriately chosen target function representing free field propagation. The matrix of prefilters $\boldsymbol{G}(k)$ generates the P loudspeaker signals $\boldsymbol{x}(k)$ from

$$\boldsymbol{x}(k) = \boldsymbol{G}(k) * \boldsymbol{x}'(k). \tag{3}$$

4. MICROPHONE ARRAY FOR STEERABLE BEAMFORMING

In a real-life environment, the signal to be recorded is subject to several more disturbances, apart from the interfering loudspeaker signals (*Fig.* 1): The reverberation of desired signals, background noise and/or competing sources (e.g., speakers). An effective approach to address these problems is to replace the single microphone by a microphone array directing a beam of increased sensitivity at the active talker, e.g. [5].

Thereby we assume a filter and sum beamformer with arbitrary filters $b_q(k)$, $q = 1, \ldots, Q$ for the microphone signals. All the known fixed and adaptive beamforming approaches can be derived from this structure.

In order to capture multiple sources simultaneously, it is possible to allow multiple beams covering all directions of interest, i.e., in general, we apply Q' beamformers to the Q microphone signals ($1 \leq Q' < Q$), denoted by vector $\boldsymbol{y}(k)$ [10]. This can be conveniently written as

$$\boldsymbol{y}'(k) = \boldsymbol{B}(k) * \boldsymbol{y}(k), \tag{4}$$

where $\boldsymbol{B}(k)$ denotes the matrix of filter impulse responses, and $\boldsymbol{y}'(k)$ is the vector of beamformer output signals.

To facilitate the integration of AEC into the microphone path, a decomposition of $\boldsymbol{B}(k)$ may be carried out, e.g., as proposed in [11]. At first, a set of Q' fixed beams is generated from the Q microphone signals. These fixed beams cover all potential sources of interest and correspond to a time-invariant impulse response

matrix $\boldsymbol{B}(k)$ in Eq. (4). The fixed beamformer is followed by a time-variant stage $\boldsymbol{V}(k)$ (voting).

The advantage of this decomposition is twofold. At first, it allows integration of MC AEC as explained below. Secondly, automatic beam steering towards sources of interest is possible, whereby external information on the positions via audio, video or multimodal object localization can be easily incorporated.

5. STRATEGIES FOR SYSTEM INTEGRATION

The concepts of wave field synthesis and steerable beamforming presented above are now set into a common framework. *Figure* 4 shows a multichannel loudspeaker-room-microphone (LRM) setup which acts as transmission and receiving room simultaneously. The loudspeaker path on top consists of P'' source signals $\boldsymbol{x}''(k)$, the aurealization matrix $\boldsymbol{A}(k)$ and the equalization prefilter $\boldsymbol{G}(k)$ which provides the P loudspeaker signals $\boldsymbol{x}(k)$. The impulse response matrix from the WFS array to the possible listener positions is given by $\boldsymbol{H}_L$, while the corresponding matrix from the WFS array to the microphone array is given by $\boldsymbol{H}(k)$. The matrices $\boldsymbol{B}(k)$ and $\boldsymbol{V}(k)$ describe the time-invariant and the time-variant components of the beamformer. Its output are the Q'' source signals $\boldsymbol{y}''(k)$ which may act as input to a WFS system at the side of the receiver.

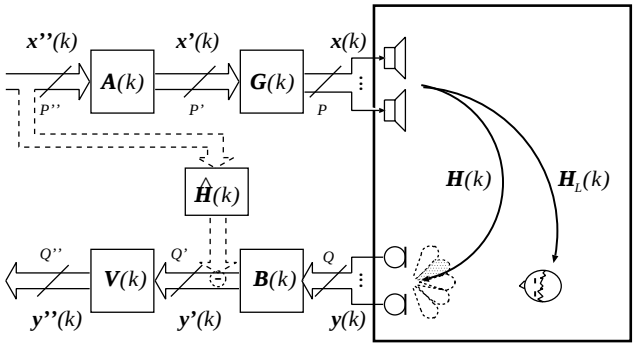

Fig. 4. Setup in terms of matrices $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{H}$

When placing the AEC between the two branches in *Fig.* 4, ideally, two objectives should be fulfilled:

- a low number of impulse responses to be identified
- a time-invariant (or very slowly time-varying) echo path

Placing the AEC in parallel to the room echoes $\boldsymbol{H}(k)$ (i.e., between $\boldsymbol{x}$ and $\boldsymbol{y}$) is prohibitive due to the high number of $P \cdot Q$ impulse responses. On the other hand, positioning the AEC between $\boldsymbol{x}''$ and $\boldsymbol{y}''$ ($P'' \cdot Q''$ impulse responses) would include the time-variant matrix $\boldsymbol{V}(k)$ into the LRM model (see [11] for a detailed discussion). Thus the most practical solution is placing the AEC between $\boldsymbol{x}''$ and $\boldsymbol{y}'$ as shown in *Fig.* 4.

Other than in the network applications considered in [12], the source signals $\boldsymbol{x}''(k)$ may be highly cross-correlated. Furthermore, to preserve a high sound quality, no aggressive decorrelation measures [5] are allowed. Section 6 describes an efficient MIMO AEC system $\hat{\boldsymbol{H}}(k)$ which can deal with correlated input signals by taking the cross-correlations explicitly into account.

Another point to consider is the total length $L_{\hat{H}}$ of the impulse responses in $\hat{\boldsymbol{H}}(k)$. Without room equalization (i.e., $\boldsymbol{G}$: identity matrix), this length is given by $L_{\hat{H}} = L_A + L_H + L_B$. With an effective room equalization, this length may actually decrease. However, even for an ideal compensation (see Eq. (2)), there will always be a direct path from $\boldsymbol{x}$ to $\boldsymbol{y}$ thus requiring AEC. These considerations show the intricate relations between the various impulse responses in *Fig.* 4.

6. MIMO-SYSTEM ACOUSTIC ECHO CANCELLATION

The performance of MC AEC is more severely affected by the choice of the adaptation algorithm than its single-channel counterpart. This is due to the very ill-conditioned nature of the underlying normal equation of the optimization problem to be solved iteratively.

For such applications, the recursive least-squares (RLS) algorithm is known to be the optimum choice in terms of convergence speed as it exhibits properties that are independent of the eigenvalue spread. The update equation of the RLS algorithm for adaptive MIMO systems reads

$$\tilde{\boldsymbol{H}}(k) = \tilde{\boldsymbol{H}}(k-1) + \boldsymbol{k}(k)\boldsymbol{e}^T(k), \quad (5)$$

$$\boldsymbol{k}(k) = \boldsymbol{R}_{xx}^{-1}\tilde{\boldsymbol{x}}''(k), \quad (6)$$

where $\tilde{\boldsymbol{H}}(k)$ denotes $(P''L_{\hat{H}} \times Q')$-matrices of adaptive MIMO filter coefficients, $\boldsymbol{e}(k) = \boldsymbol{y}'(k) - \hat{\boldsymbol{y}}'(k)$ is the current residual error vector of length-Q' between the echoes and the echo replicas, and $\boldsymbol{k}(k)$ is the so-called Kalman gain vector. The calculation of $\boldsymbol{k}(k)$ is the computationally most demanding part due to the inversion and multiplication of the large $(P''L_{\hat{H}} \times P''L_{\hat{H}})$ correlation matrix $\boldsymbol{R}_{xx}$. The length-$P''L_{\hat{H}}$ vector $\tilde{\boldsymbol{x}}''(k)$ is a concatenation of the input signal vectors containing the $L_{\hat{H}}$ most recent input samples. Note that $\boldsymbol{R}_{xx}$ contains both auto-correlations and all cross-correlations between the input channels. The major problems of RLS or fast RLS algorithms are the very high level of computational complexity and potential numerical instabilities making it difficult to implement for real-time applications.

The adaptive filters in our real-time system are efficiently updated in the frequency domain in a block-by-block fashion, using the fast Fourier transform (FFT) as a powerful vehicle. The block length can be chosen as $N \leq L_{\hat{H}}$, where $N = L_{\hat{H}}$ is most efficient. As a result of this block processing, the arithmetic complexity is significantly reduced compared to time-domain adaptive algorithms while desirable RLS-like properties and the basic structure of Eq. (5) and Eq. (6) are maintained in our algorithm.

The possibility to exploit the efficiency of FFT algorithms is due to the Toeplitz structure of the matrices involved, which results from the time-shift properties of the input signals. Consequently, by rewriting the original time-domain optimization criterion in a way that Toeplitz and circulant matrices are explicitly shown allows a mathematically rigorous derivation of single- and multichannel frequency-domain adaptive algorithms, as shown in [1, 4]. Due to the rigorous approach, the framework inherently takes the correlations between all input channels into account but approximately diagonalizes the correlation matrices. *Figure* 5 gives an overview of the scheme. Note that the calculation of the *frequency-domain Kalman gain* for MIMO systems is independent of the number of beam signals in $\boldsymbol{y}'(k)$ allowing significant computational savings [2]. For efficient and stable methods to calculate it, see [3, 4].

7. SIMULATIONS AND REAL-TIME IMPLEMENTATION

Our real-time implementation of the whole system runs on a regular PC platform with multichannel sound card and a loudspeaker

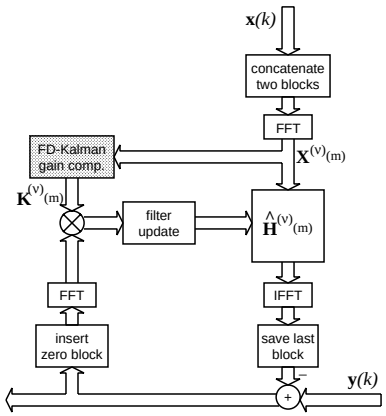

Fig. 5. Frequency-domain adaptive MIMO filter.

array with 24 elements and a nested microphone array with 11 elements. The transducer setup can easily be extended. For the echo cancellation it is possible to trade signal bandwidth and filter lengths (4096 per channel is reasonable on current PCs) with the number of input sources ($1 \leq P'' \leq 5$).

For the following AEC simulation curves, a speech signal (in the transmission room) was convolved by P'' different room impulse responses and nonlinearly, but inaudibly preprocessed according to [5] (P'' different nonlinearities with factor 0.5). The lengths of the receiving room impulse responses were 4096. To allow a fair comparison with other AEC approaches, independent white noise signals for an echo-to-noise-ratio of $35dB$ were added to the echo on each microphone. *Fig.* 6 shows the misalignment convergence of the described algorithm (solid) for the multi-channel cases $P'' = 2, 3, 4, 5$ (from lowest to uppermost line). The input signal blocks were overlapped by a factor $\alpha = 4$, i.e., each block contains 25% previously unseen samples. In *Fig.* 7 the overlap factor α was adjusted to 8 for $P'' = 3, 4$ and to 16 for $P'' = 5$. One can clearly see that the achieved performance is then nearly equal for all numbers of channels P'' considered here. The dashed lines show the corresponding characteristics for the basic NLMS algorithm that does not take cross-correlations into account, e.g. [5].

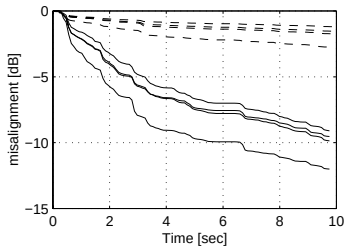

Fig. 6. Convergence for P''=2,3,4,5 channels, $\alpha = 4$

8. CONCLUSIONS

Recent progress in the field of multichannel acoustic echo cancellation motivates the combination with new 3D sound reproduction techniques such as the wave field synthesis, and multichannel sound acquisition for beamforming. From the single components of such a system, we established a matrix framework and discussed

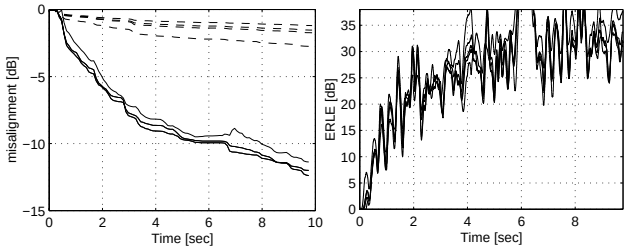

Fig. 7. Convergence for for P''=2,3,4,5 channels and adjusted overlap α

different options. While a completely general solution is still very challenging, a suboptimum and efficient solution that is interesting for many applications has been presented. Our first real-time implementation runs on a regular PC platform.

9. REFERENCES

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

[2] H. Buchner, W. Herbordt, and W. Kellermann,"An efficient combination of multichannel acoustic echo cancellation with a beamforming microphone array," in *Conf. Rec. Int. Workshop on Hands-Free Speech Commun.*, pp. 55-58, April 2001.

[3] H. Buchner and W. Kellermann,"Acoustic echo cancellation for two and more reproduction channels," in *Conf. Rec. IEEE IWAENC*, pp. 99-102, Sept. 2001.

[4] H. Buchner, J. Benesty, and W. Kellermann,"Generalized multichannel frequency-domain adaptive filtering: efficient realization and application to hands-free speech communication," *IEEE Trans on SP*, submitted, Feb. 2002.

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

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

[7] A.J. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wave field synthesis," *Journal of the Acoustic Society of America*, vol. 93, no. 5, pp. 2764–2778, May 1993.

[8] D. de Vries, E. Hulsebos, and E. Bourdillat, "Improved microphone array configurations for auralization of sound fields by Wave Field Synthesis," *110th Conv. of the AES*, May 2001.

[9] U. Horbach et al., "Numerical simulation of wave fields created by loudspeaker arrays," *107th Conv. of the AES*, Sept. 1999.

[10] M. S. Brandstein and D. B. Ward (eds.), *Microphone Arrays: Techniques and Applications*, Springer Verlag, Berlin, 2001.

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

[12] T. Yensen et al., "Synthetic stereo acoustic echo cancellation structure with microphone array beamforming for VoIP conferences," in *Proc. ICASSP 2000*, pp. 817-820.