FULL-DUPLEX MULTICHANNEL COMMUNICATION: REAL-TIME IMPLEMENTATIONS IN A GENERAL FRAMEWORK

W. Herbordt, H. Buchner, W. Kellermann, R. Rabenstein, S. Spors, and H. Teutsch

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

ABSTRACT

In this contribution, we embed full-duplex multichannel communication interfaces for tele-presence systems into a general framework. On the reproduction side, we consider a wide range of multichannel acoustic rendering techniques including traditional stereophony, '5.1' systems, and wave field synthesis using loudspeaker arrays for sound immersion. On the recording side, microphone arrays are discussed for capturing clean desired signals with spatial information. Based on this general framework, real-time implementations of such full-duplex multichannel communication systems are then described. We combine wave field synthesis with multichannel acoustic echo cancellation and adaptive beamforming and discuss a real-time implementation on standard desktop and laptop PCs.

1. INTRODUCTION

We consider the full-duplex acoustic human-machine interface after Figure 1. On the reproduction side, sound effects are produced by stereophonic or '5.1' setups with present commercial sound systems. For reproducing acoustic scenes more faithfully, wave field synthesis recreates 3D acoustic wave fields with a large number of loudspeakers (tens or hundreds) based on the physical description of the original acoustic environment using Huygens' principle [1]. On the recording side, signals of desired local sources and undesired local noise are captured by microphone arrays. Compared to single-channel recording, microphone arrays allow to preserve the spatial sound impression and to filter the local wave field spatially. Ideally, the users should be allowed to move freely within the acoustic environment, which usually means that loudspeakers and microphones are not positioned close to the users and that relative positions vary. This general scenario covers most hands-free full-duplex applications for communication in natural or virtual acoustic environments as, e.g., cinemas, home theaters, vehicle interiors, tele-conferencing, tele-presence, teleteaching, multimedia terminals, speech dialog systems, or virtual reality environments.

This hands-free scenario may be efficiently described as follows: The acoustic rendering part receives K source signals captured in a vector $\mathbf{u}$ and renders them using L loudspeaker signals $\mathbf{v}$. The M local users capture $2M$ signals $\mathbf{w}$ with their ears, where $\mathbf{w}$ consists of the reproduced loudspeaker signals plus additive local noise $\mathbf{n_w}$. Describing the room impulse responses between the loudspeakers and the ears by multiple-input / multiple-output (MIMO) system $\mathbf{H_{wv}}$, we can write the listeners' signals $\mathbf{w}$ as [1]:

$$\mathbf{w} = \mathbf{H_{wv}} * \mathbf{v} + \mathbf{n_w} \,. \tag{1}$$

The N microphone signals $\mathbf{x}$ of the recording part contain the M desired source signals s_i of the local users S_i, which are summarized in a vector $\mathbf{s}$, echoes of the loudspeaker signals $\mathbf{v}$, and additive local noise $\mathbf{n_x}$:

$$\mathbf{x} = \mathbf{H_{xs}} * \mathbf{s} + \mathbf{H_{xv}} * \mathbf{v} + \mathbf{n_x} \,. \tag{2}$$

$\mathbf{H_{xv}}$ and $\mathbf{H_{xs}}$ capture the room impulse responses between the corresponding vector components in MIMO systems.

Note that for reproduction, the MIMO system $\mathbf{H_{wv}}$ describes the transfer characteristic to any point in the listening area. Obviously, stereo and '5.1' system only account for a small subspace ('sweet spot') whereas wave field synthesis optimizes $\mathbf{H_{wv}}$ for the entire listening area.

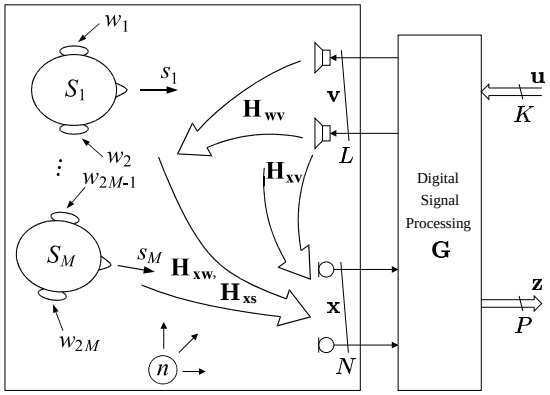

Fig. 1. *Full-duplex communication system.*

This work was partly supported by grants from Intel Corp. (Beijing, China; Santa Clara, CA), from Grundig AG, Nuremberg, leading the German EMBASSI consortium, and from the European Commission as sponsor of the CARROUSO project.

[1] The linear convolution $\mathbf{y} = \mathbf{A} * \mathbf{x}$ between a column vector $\mathbf{x}$ with elements $x_i(k)$ and a matrix $\mathbf{A}$ with time-invariant impulse responses $a_{ij}(k)$ is defined by $y_i(k) = \sum_{j=1}^{N} \sum_{n=-\infty}^{\infty} a_{ij}(k-n) x_j(n)$, where $y_i(k)$ is the i-th component of $\mathbf{y}$.

Using this general framework, we first identify the components which are required for full-duplex communication systems (Section 2). Second, we present a strategy for integrating wave field synthesis, multichannel sound recording, and multichannel acoustic echo cancellation into a real-time system in Section 3. Finally, in Section 4, we outline our present real-time hardware setup.

2. AUDIO SIGNAL PROCESSING FOR FULL-DUPLEX COMMUNICATION SYSTEMS

Ideally, the full-duplex communication system $\mathbf{G}$ processes the source signals $\mathbf{u}$ and the sensor signals $\mathbf{x}$ such that $\mathbf{w}$ corresponds to a desired sound impression $\mathbf{w}_\mathrm{d}$ and such that the output vector $\mathbf{z}$ consists of $P \leq M$ desired signals, respectively.

Assuming that the matrices $\mathbf{H_{wv}}$, $\mathbf{H_{xv}}$, and $\mathbf{H_{xs}}$ can be modeled as linear discrete-time systems, $\mathbf{G}$ is linear and the signal processing can be completely described by linear convolutions as

$$\begin{pmatrix} \mathbf{v} \\ \mathbf{z} \end{pmatrix} = \mathbf{G} * \begin{pmatrix} \mathbf{u} \\ \mathbf{x} \end{pmatrix} = \begin{pmatrix} \mathbf{G_{vu}} & \mathbf{G_{vx}} \\ \mathbf{G_{zu}} & \mathbf{G_{zx}} \end{pmatrix} * \begin{pmatrix} \mathbf{u} \\ \mathbf{x} \end{pmatrix}, \quad (3)$$

where the submatrices $\mathbf{G_{vu}}$, $\mathbf{G_{vx}}$, $\mathbf{G_{zu}}$, and $\mathbf{G_{zx}}$ describe the signal processing between the indexed signal vectors.

Using this matrix description, we can derive conditions which need to be fulfilled by $\mathbf{G}$. We assume that the signals $\mathbf{s}$ and $\mathbf{u}$ are mutually uncorrelated and uncorrelated with $\mathbf{n_w}$, $\mathbf{n_x}$.

2.1. Multichannel sound reproduction

The acoustic rendering part processes the source signals $\mathbf{u}$ such that an acoustic impression $\mathbf{w}_\mathrm{d}$ is produced, i.e.:

$$\mathbf{w} \stackrel{!}{=} \mathbf{w}_\mathrm{d} = \mathbf{H}_\mathrm{d} * \mathbf{u}\,, \quad (4)$$

where the $2M \times K$-matrix $\mathbf{H}_\mathrm{d}$ contains the generally time-varying desired impulse responses $h_{ij}(k,l)$ between input channel u_j and ear i. Using (1) and (3), we obtain

$$\mathbf{H_{wv}} * (\mathbf{G_{vu}} * \mathbf{u} + \mathbf{G_{vx}} * \mathbf{x}) + \mathbf{n_w} \stackrel{!}{=} \mathbf{H}_\mathrm{d} * \mathbf{u}\,. \quad (5)$$

In order to fulfill this condition, two signal processing techniques are involved:

2.1.1. Deconvolution

The matrix $\mathbf{G_{vu}}$ must invert the influence of the room impulse responses $\mathbf{H_{wv}}$ if the signal processing for sound reproduction should be independent of $\mathbf{u}$: [2]

$$\begin{aligned} \mathbf{H_{wv}} * \mathbf{G_{vu}} * \mathbf{u} &\stackrel{!}{=} \mathbf{H}_\mathrm{d} * \mathbf{u} \\ \rightarrow \mathbf{H_{vw}} * \mathbf{G_{uv}} &\stackrel{!}{=} \mathbf{H}_\mathrm{d} \\ \rightarrow \mathbf{G_{vu}} &\stackrel{!}{=} \mathbf{H_{wv}^{-1}} * \mathbf{H}_\mathrm{d}\,. \end{aligned} \quad (6)$$

Obviously, we need to allow for a delay $\mathbf{H}_\mathrm{d}$ in order to assure causality of $\mathbf{G_{vu}}$. For inversion of the impulse responses $\mathbf{H_{wv}}$, we first need in general to identify $\mathbf{H_{wv}}$. For conventional sound reproduction systems as, e.g., stereophony or '5.1' systems, this means that $\mathbf{H_{wv}}$ must be identified for the present positions of the users, which are generally strongly time-varying. We obtain a blind deconvolution problem with unknown output signal. For identification of the current $\mathbf{H_{wv}}$, the signals at the ears are required, which are generally not available. For loudspeaker arrays using wave field synthesis, ideally, the *entire* listening room without the influence of local users is compensated [2]. This is desirable for scenarios, where the user should have the impression to be an active part of the reproduced scene as in tele-presence, or virtual reality environments. For scenarios where the user should have a predefined sound impression such as cinemas or home theaters it is necessary to compensate the influence of both listening room and local users.

[2] The inverse matrix $\mathbf{A}^{-1}$ is defined as the matrix which fulfills $\mathbf{A}^{-1} * \mathbf{A} = \mathbf{I} \cdot \delta(k)$, where $\mathbf{I}$ is the identity matrix and where $\delta(k)$ is the discrete-time unit impulse. If $\mathbf{A}$ is not invertible, then, $\mathbf{A}^{-1}$ denotes the pseudoinverse of $\mathbf{A}$.

2.1.2. Noise compensation

Local noise is cancelled at the ears of the users if a noise compensation signal is fed to the loudspeakers such that the condition

$$\begin{aligned} \mathbf{H_{wv}} * \mathbf{G_{vx}} * \mathbf{x} + \mathbf{n_w} &\stackrel{!}{=} \mathbf{0} \\ \rightarrow \mathbf{G_{vx}} * \mathbf{x} &\stackrel{!}{=} -\mathbf{H_{wv}^{-1}} * \mathbf{n_w} \end{aligned} \quad (7)$$

is solved. We identify three principal tasks for compensation of $\mathbf{n_w}$. First, a reference signal for the noise $\mathbf{n_w}$ at the ears is required, which is given by $\mathbf{H_{xw}^{-1}} * \mathbf{n_x}$. $\mathbf{G_{vx}}$ thus must extract $\mathbf{n_x}$ from the sensor signals, which are a mixture of $\mathbf{s}$, $\mathbf{v}$, and $\mathbf{n_x}$. Second, the inverse of $\mathbf{H_{xw}}$ must be determined. This corresponds to blind deconvolution with unknown input signals $\mathbf{n_w}$. Third, $\mathbf{G_{vx}}$ must inversely model $\mathbf{H_{wv}}$. The inversion of $\mathbf{H_{wv}}$ was identified in Section 2.1.1 as a blind deconvolution problem with unknown output signals. Note that (7) formally corresponds to the active noise cancellation problem [3], which is only resolved for few applications so far.

2.2. Multichannel sound recording

Multichannel sound recording aims at the extraction of P desired signals, which requires suppression of local noise and cancellation of acoustic echoes of the loudspeaker signals. The desired signals are generally convolutive mixtures of the desired source signals $\mathbf{s}$:

$$\begin{aligned} \mathbf{z} &= \mathbf{G_{zu}} * (\mathbf{u} + \mathbf{G_{zx}} * \mathbf{x}) \\ &= \mathbf{G_{zu}} * \mathbf{u} + \mathbf{G_{zx}} * (\mathbf{H_{xs}} * \mathbf{s} + \mathbf{H_{xv}} * \mathbf{v} + \mathbf{n_x}) \\ &= (\mathbf{G_{zu}} + \mathbf{G_{zx}} * \mathbf{H_{xv}} * \mathbf{G_{vu}}) * \mathbf{u} \\ &\quad + \mathbf{G_{zx}} * (\mathbf{H_{xs}} * \mathbf{s} + \mathbf{n_x}) \\ &\stackrel{!}{=} \mathbf{H_{zs}} * \mathbf{s}\,. \end{aligned} \quad (8)$$

Hence, we can relate three signal processing techniques to sound recording.

2.2.1. Acoustic echo cancellation

In order to compensate the feedback between the signals $\mathbf{v}$ and the microphone signals $\mathbf{x}$, the condition

$$(\mathbf{G_{zu}} + \mathbf{G_{zx}} * \mathbf{H_{xv}} * \mathbf{G_{vu}}) * \mathbf{u} \stackrel{!}{=} \mathbf{0} \quad (9)$$

must be met. If $\mathbf{G_{zx}}$ should be independent of the acoustic echoes and if the acoustic echoes should be canceled independently of $\mathbf{v}$

$$\mathbf{G_{zu}} \stackrel{!}{=} -\mathbf{G_{zx}} * \mathbf{H_{xv}} * \mathbf{G_{vu}} \tag{10}$$

must be met. Acoustic echo cancellation is thus equivalent to multichannel system identification, where input and output signals can be observed and where the matrix $\mathbf{H_{xv}}$ of impulse responses between loudspeaker signals and microphone signals has to be identified [4]. The system $\mathbf{G_{zu}}$ between the signals $\mathbf{u}$ and $\mathbf{z}$ must be equivalent to the feedback path which consists of the cascade $\mathbf{G_{zx}}$, $\mathbf{H_{xv}}$, and $\mathbf{G_{vu}}$. This is corresponds to placing the acoustic echo cancellation directly between $\mathbf{u}$ and $\mathbf{z}$. However, acoustic echoes are equivalently suppressed, if the echo canceller is placed between the loudspeaker signals $\mathbf{v}$ and the sensor signals $\mathbf{x}$, which means that only $\mathbf{H_{xv}}$ must be modelled. This might be computationally less efficient, which depends on the number of echo paths. The identification of the echo paths $\mathbf{H_{xv}}$ is still difficult: First, the echo components in $\mathbf{x}$ are usually not available separately. Second, the reference signals $\mathbf{u}$ are generally highly auto-correlated and highly cross-correlated. Third, acoustic echo paths generally have impulse responses with large number of filter taps (hundreds to thousands), which makes computationally efficient convolution necessary.

2.2.2. Noise reduction

For suppression of local noise, the condition

$$\mathbf{G_{zx}} * \mathbf{n_x} \stackrel{!}{=} \mathbf{0} \tag{11}$$

must be fulfilled. Obviously, the solution is signal-dependent. If $\mathbf{G_{zx}}$ should be independent of $\mathbf{n_x}$, $\mathbf{G_{zx}} = \mathbf{0}$ is required. This, however, prevents recording of the desired signals. $\mathbf{G_{zx}}$ may be realized as multichannel noise reduction, which requires estimates of $\mathbf{n_x}$ or as a beamformer, which requires knowledge of spatial characteristics of the noise [5].

2.2.3. Source separation and dereverberation

Finally, the sensor signals need to be processed such that

$$\mathbf{G_{zx}} * \mathbf{H_{xs}} * \mathbf{s} \stackrel{!}{=} \mathbf{H_{zs}} * \mathbf{s}\,, \tag{12}$$

which yields for signal-independent solutions the condition

$$\mathbf{G_{zx}} = \mathbf{H_{zs}} * \mathbf{H_{xs}^{-1}}\,. \tag{13}$$

We thus face a multichannel blind inversion problem ('dereverberation') and a multichannel noise reduction problem for the elements on the main diagonal and for the elements on the off-diagonals of $\mathbf{G_{zx}} * \mathbf{H_{xs}}$, respectively. Convolution with $\mathbf{H_{zs}}$ assures causality and adds the desired spatial sound impression to the separated desired signals $\mathbf{x}$. In the simplest case, $\mathbf{H_{zs}}$ is a diagonal matrix with delayed unit impulses on the main diagonal.

Compared to single-channel techniques, which allow separation in the time domain and in the frequency domain, multichannel recordings enable exploitation of spatial information: With beamforming, coherent signals are separated based on directions of incidence. Beamformers have to cope with non-stationarity and reverberation. In this case, adaptive beamformers are promising [5]. Blind source separation extracts coherent signals based on statistical methods which exploit statistical independence of the sources or uncorrelatedness in conjunction with non-stationarity and non-Gaussianity assumptions [6, 7].

3. SYSTEM INTEGRATION

We show now how we can combine above components for multichannel sound reproduction and multichannel sound recording for investigating full-duplex communication systems in real-time (see Figure 2, based on [8]). For high-quality acoustic rendering, we use wave field synthesis with a loudspeaker array. Virtual room acoustics are created by designing the auralization matrix $\mathbf{H_d}$ [1]. As mentioned in Section 2.1.1, wave field synthesis allows to compensate the influence of the listening environment with the matrix $\mathbf{G_{vu}}$ [2].

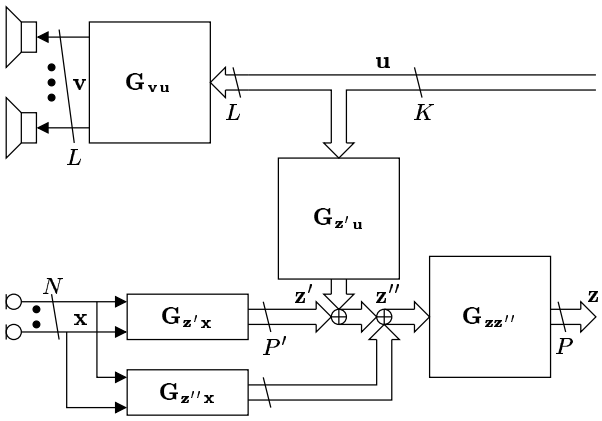

Fig. 2. *Real-time implementation of a multichannel full-duplex communication system.*

On the recording side, a microphone array picks up the mixture of the signals of the local users, the echoes of loudspeaker signals, and local noise. The acoustic echo canceller $\mathbf{G_{z'u}}$ uses $\mathbf{u}$ as reference and acts on internal signals of $\mathbf{G_{zx}}$. The post-processing of the sensor signals for dereverberation, for noise reduction, and for signal separation is separated into three blocks $\mathbf{G_{z'u}}$, $\mathbf{G_{z''x}}$, and $\mathbf{G_{zz''}}$. The matrix $\mathbf{G_{z'x}}$ captures P' fixed beamformers, which are steered to P' possible positions of the local users. $\mathbf{G_{z'x}}$ reduces the noise, partially dereverberates, and partially separates the signals in the steering directions [5]. The matrix $\mathbf{G_{z''x}}$ consists of P' adaptive beamformers between $\mathbf{x}$ and $\mathbf{z}'$ which are required for efficient separation of transient desired signals and transient noise. Note that this structure corresponds to P' parallel generalized sidelobe cancellers [9], which are steered to P' positions.

By placing the multichannel acoustic echo canceller between $\mathbf{u}$ and $\mathbf{z}'$, we minimize computational complexity by minimizing the number of echo paths (i.e. $K < L$, $P' < N$) and we assure that the identification of the echo paths is not complicated by the time-variance of the adaptive beamformers $\mathbf{G_{z''x}}$ [10]. The combination of generalized sidelobe canceller and multichannel acoustic echo canceller is implemented as described in [11]. The matrix $\mathbf{G_{zz''}}$ forms the output signals of the communication system from

the P' spatially selective signals $\mathbf{z}''$. It selects desired signals using a voting algorithm [12] and adds spatial information about the desired signals if multichannel reproduction is desired.

4. REAL-TIME SETUP

Our real-time setup of the full-duplex multichannel communication system is depicted in Figure 3. It consists of up to 24 wideband loudspeakers, a subwoofer, and a line microphone array with up to 26 logarithmically spaced elements [8, 13]. The transducers are connected to multichannel soundcards on 2 regular dual-processor PC platforms - one for the reproduction side and one for the recording side. The real-time software was developed partly for LINUX and partly for Microsoft Windows operating systems using BruteFIR libraries [14] for fast convolution on the reproduction side and Intel Integrated Performance Primitives libraries [15] for efficient signal processing on the recording side.

Fig. 3. *24-channel loudspeaker array and 26-channel microphone array in a multimodal full-duplex communication interface.*

5. CONCLUSIONS

We described full-duplex multichannel communication systems in a general system theoretical framework using linear matrix formulations. The various signal processing tasks can be summarized as signal separation and system identification with different challenges, which are not completely resolved so far. We showed, however, that many of the problems of full-duplex communication systems can be resolved: Loudspeaker arrays with wave field synthesis allow for high-quality acoustic rendering including compensation of the listening room. Microphone arrays combined with adaptive beamforming and multichannel acoustic echo cancellation enable high-quality sound recording even in adverse acoustic conditions. Our real-time implementation efficiently combines these components such that high-quality full-duplex communication can now be realized on standard PC hardware in real-time.

6. REFERENCES

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

[2] S. Spors, A. Kuntz, and R. Rabenstein, "Listening room compensation for wave field synthesis," *(accepted to) IEEE Int. Conf. on Multimedia and Expo*, July 2003.

[3] S.M. Kuo and D.R. Morgan, *Active noise control systems*, John Wiley & Sons, Inc., New York, 1996.

[4] H. Buchner, J. Benesty, and W. Kellermann, "Multichannel frequency-domain adaptive filtering with application to multichannel acoustic echo cancellation," in *Adaptive Signal Processing: Applications to Real-World Problems*, J. Benesty and Y. Huang, Eds. Springer, Berlin, 2003.

[5] W. Herbordt and W. Kellermann, "Adaptive beamforming for audio signal acquisition," in *Adaptive Signal Processing: Applications to Real-World Problems*, J. Benesty and Y. Huang, Eds., pp. 155–194. Springer, Berlin, 2003.

[6] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of a class of blind source separation algorithms for convolutive mixtures," *Proc. Fourth Int. Symposium on Independent Component Analysis and Blind Source Separation*, April 2003.

[7] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, Inc., New York, 2001.

[8] H. Buchner, S. Spors, W. Kellermann, and R. Rabenstein, "Full-duplex communication systems with loudspeaker arrays and microphone arrays," *IEEE Proc. Int. Conf. on Multimedia and Expo*, August 2002.

[9] H.L. Van Trees, *Optimum Array Processing, Part IV of Detection, Estimation, and Modulation Theory*, John Wiley, New York, 2002.

[10] W. Kellermann, "Acoustic echo cancellation for beamforming microphone arrays," in *Microphone Arrays: Signal Processing Techniques and Applications*, M.S. Brandstein and D.B. Ward, Eds., chapter 13, pp. 281–306. Springer, Berlin, 2001.

[11] W. Herbordt, H. Buchner, and W. Kellermann, "An acoustic human-machine front-end for multimedia applications," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 1, pp. 1–11, January 2003.

[12] W. Kellermann, "A self-steering digital microphone array," *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 5, pp. 3581–3584, May 1991.

[13] S. Spors, H. Teutsch, and R. Rabenstein, "High-quality acoustic rendering with wave field synthesis," *Vision, Modeling, and Visualization*, pp. 101–108, November 2002.

[14] A. Torger, "Brutefir - an open-source general-purpose audio convolver," http://www.ludd.luth.se/ torger/brutefir.html.

[15] Intel Corp., "Intel integrated performance primitives 3.0," http://developer.intel.com/software/products/ipp/ipp30/.