ACOUSTIC SIGNAL PROCESSING FOR NEXT-GENERATION HUMAN/MACHINE INTERFACES

Walter Kellermann

Chair for Multimedia Communications and Signal Processing
University Erlangen-Nuremberg, Germany
wk@LNT.de

ABSTRACT

In this paper, we first define the scenario of a generic acoustic human/machine interface and then formulate the according fundamental signal processing problems. For signal reproduction, the requirements for ideal solutions are stated and some examples for the state of the technology are briefly reviewed. For signal acquisition, the fundamental problems ask for acoustic echo cancellation, desired source extraction, and source localization. After illustrating to which extent acoustic echo cancellation is already a solved problem, we present recent results for separation, dereverberation and localization of multiple source signals. As an underlying motivation for this synoptic treatment, we demonstrate that the considered subproblems (except localization) can be directly interpreted as signal separation or system identification problems with varying degrees of difficulty, which in turn determines the effectiveness of the known solutions.

1. INTRODUCTION

Over the last century human/machine interaction became a more and more common part of our everyday life and along with increasingly complex machines the demand for more 'human' interfaces grew continuously. With speech still being the most efficient modality for communication involving humans, and audio being an ubiquitously desired commodity, the acoustic component very often plays a dominant role in the design of human/machine interfaces. For many situations, the ideal 'natural' acoustic human/machine interface should allow the users to be untethered, mobile and distant from the signal acquisition and reproduction equipment without the need to wear any extra devices. Such a situation is illustrated in Fig.1 where we consider several users in an acoustic environment with multichannel sound reproduction and a microphone array for multichannel audio acquisition. On the reproduction side, vector $\mathbf{v}$ contains L loudspeaker signals, which are derived from (or identical with) the vector of K source signals $\mathbf{u}$. Vector $\mathbf{w}$ describes the $2M$ signals at the ears of the M listeners, which in the ideal case correspond to a set of desired signals $\mathbf{w}_\mathrm{d}$. Correspondingly, $\mathbf{s}$ represents the signals emitted by M desired sources S_i, which should be captured by N microphones. Vector $\mathbf{n}$ accounts for any unwanted acoustic signals, whose contributions to the microphone signals and the ear signals are termed $\mathbf{x_n}$, $\mathbf{w_n}$, respectively. From the vector of N microphone signals $\mathbf{x}$, the acquisition part of the digital signal processing unit $\mathbf{G}$ extracts a vector $\mathbf{z}$, whose elements are ideally identical to $P \leq M$ desired signals s_i. The matrices $\mathbf{H_{wv}}$,$\mathbf{H_{xv}}$, $\mathbf{H_{xs}}$ describe the transfer characteristics between the respective vectors. In this general scenario, the tasks for the digital signal processing (DSP) unit $\mathbf{G}$ are

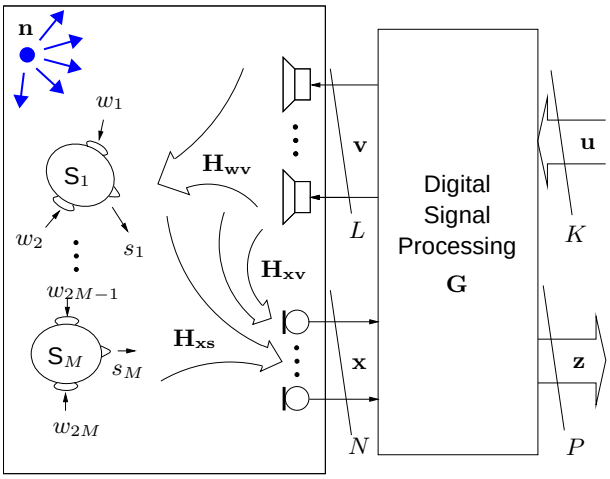

Figure 1: *Multichannel acoustic human/machine interface.*

simply stated:

- The reproduction part should deliver well-defined desired signals $\mathbf{w}_\mathrm{d}$ to the listeners' ears.
- The signal acquisition should extract the desired source signals $\mathbf{s}$ and determine the source locations.

In realistic environments these tasks involve obviously three sources of problems, which have to be addressed by the DSP algorithms: Noise, echoes and reverberation, and echoes of loudspeaker signals feeding back into the microphones.

The setup in Fig.1 covers a multitude of applications, with varying emphasis and difficulty regarding the different problems of the general scenario. A wide class of applications can be summarized as hands-free equipment for telecommunication and human/machine dialogues involving speech recognition systems: For mobile phones, personal digital assistants and mobile computing devices, hands-free operation is becoming increasingly popular. In cars, hands-free equipment is now often an integral part of the user front-end for telephony, infrastructure control, and navigation systems. Seamless voice interaction with desktop computers, multimedia terminals, and game stations is another field of applications with a large market potential. Teleconferencing equipment, now ranging from desktop computer accessories over especially equipped rooms to large auditoria has been one of the initial applications [1, 2]. Telecollaboration and teleteaching systems may be viewed as belonging to the same category. Another

group of hands-free applications with even greater emphasis on human/machine voice dialogue includes smart homes, home theatre systems, smart meeting rooms, and home care for elderly people, as well as interactive museums and exhibitions.

Other areas with a stronger emphasis on the reproduction part may be summarized as audio communication applications and include cinema sound systems, equipment for virtual reality, stages and recording studios, or systems for telecollaboration of musicians.

A third and increasingly important class of applications is dedicated to acoustic surveillance, where the signal acquisition is the dominant part and the desired sources are usually non-cooperative. Here localization and signal extraction (e.g., for subsequent classification) will be the main goals of signal processing at the human/machine interface.

In the following, we first review the fundamental problems for signal processing in our scenario, thereby following partly an earlier presentation [3]. After discussing briefly the state of the art in signal reproduction, we concentrate on signal acquisition techniques and present some recent results with some bias towards work in our own research group.

2. FUNDAMENTAL SIGNAL PROCESSING PROBLEMS

For the following we assume - unless otherwise stated - that the components of the acoustic scenario can be modelled by linear, generally time-varying discrete-time systems, so that we can describe the input/output relations by matrix equations. Accordingly, the MIMO ('multiple input/multiple output') system $\mathbf{G}$ performs linear convolutions on the time-domain signals u_i, x_j ($i = 1,\ldots,K; j = 1,\ldots,N$). Decomposing $\mathbf{G}$ into submatrices $\mathbf{G_{vu}}$, $\mathbf{G_{vx}}$, $\mathbf{G_{zu}}$, $\mathbf{G_{zx}}$, we can write[1]:

$$\begin{pmatrix} \mathbf{v} \\ \mathbf{z} \end{pmatrix} = \mathbf{G} * \begin{pmatrix} \mathbf{u} \\ \mathbf{x} \end{pmatrix} = \begin{pmatrix} \mathbf{G_{vu}} & \mathbf{G_{vx}} \\ \mathbf{G_{zu}} & \mathbf{G_{zx}} \end{pmatrix} * \begin{pmatrix} \mathbf{u} \\ \mathbf{x} \end{pmatrix}. \quad (1)$$

The signals at the listeners' ears, $\mathbf{w}$, and the microphone signals, $\mathbf{x}$, are then given by

$$\mathbf{w} = \mathbf{H_{ws}} * \mathbf{s} + \mathbf{H_{wv}} * \mathbf{v} + \mathbf{w_n}, \quad (2)$$

$$\mathbf{x} = \mathbf{H_{xs}} * \mathbf{s} + \mathbf{H_{xv}} * \mathbf{v} + \mathbf{x_n}. \quad (3)$$

We emphasize here that the elements of the matrices $\mathbf{H}_{(\cdot)}$ are impulse responses which are mainly characterized by the acoustic environment with a reverberation time T_{60} [5] in the range of hundreds of milliseconds. Therefore, appropriate digital FIR filter models require several hundred to several thousand coefficients, depending on T_{60} and the sampling rate f_s. (As a rule of thumb, $L_G = f_\mathrm{s} \cdot T_{60}/3$ coefficients of the impulse response are needed for a modelling error smaller than -20dB relative to the entire impulse response energy. As an example, for a usual office and telephone signal bandwidth, $f_\mathrm{s} = 8\mathrm{kHz}$, $L_G = 1000$ is a typical choice.) Besides the mere length of the impulse responses, the according discrete-time transfer functions exhibit nonminimum phase and many zeroes close to the unit circle, which makes inversion mostly difficult and impractical. Moreover, acoustic impulse responses are strongly time-variant, not at least due to the temperature-dependency of sound velocity. Note also that $\mathbf{H_{wv}}, \mathbf{H_{ws}}$ include the head related transfer functions of the listeners.

Based on the system representation given by Eqs.1,2,3, we analyze now the fundamental problems to be solved by the signal processing unit $\mathbf{G}$. Thereby, we may safely assume that the speech signals s_i and the reproduction signals u_i (as contained in $\mathbf{v}$) are mutually statistically independent and also independent from the elements of the noise vectors $\mathbf{w_n}$, $\mathbf{x_n}$. Note also, that even if $\mathbf{G}$ connects inputs and outputs by linear filtering, $\mathbf{G}$ will usually incorporate nonlinear algorithms for determining these filters.

2.1. Sound reproduction

The goal of providing desired signals $\mathbf{w}_\mathrm{d}$ to the listeners' ears may come in two versions: In the first case, local noise should be suppressed but other desired sources S_i should still be audible in a natural way (as probable, e.g., in a teleconferencing environment). In the second case, e.g. in a cinema, the listeners will want to hear only sound resulting from the reproduction signals $\mathbf{u}$ and no interference from noise or other local sources S_i. The first case can be expressed as

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

where $\mathbf{w_s} = \mathbf{H_{ws}} * \mathbf{s}$. Introducing Eqs.1,2 leads to

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

and implies two kinds of signal processing tasks:

A. Dereverberation. Matrix $\mathbf{G_{vu}}$ has to equalize the influence of the room impulse responses $\mathbf{H_{wv}}$ on $\mathbf{u}$, and if perfectly equalized reproduction should be independent of the signal vector $\mathbf{u}$, $\mathbf{H_{wv}}$ must be inverted:

$$\mathbf{H_{wv}} * \mathbf{G_{vu}} * \mathbf{u} \stackrel{!}{=} \mathbf{H}_\mathrm{d} * \mathbf{u} \implies \mathbf{H_{wv}} * \mathbf{G_{uv}} \stackrel{!}{=} \mathbf{H}_\mathrm{d}$$
$$\implies \mathbf{G_{vu}} \stackrel{!}{=} \mathbf{H_{wv}^{-1}} * \mathbf{H}_\mathrm{d}. \quad (6)$$

Aside from a necessary delay in $\mathbf{H}_\mathrm{d}$ for assuring causality of $\mathbf{G_{vu}}$, the main problem is that $\mathbf{H_{wv}}$ is in general not known and cannot easily be identified due to lack of a reference signal at the listeners' ears (assuming, of course, that the users do not wear a microphone at the ear). Thus, we face a 'blind deconvolution' problem, where we cannot even observe the system output. A possible remedy could be to measure suitable room impulse responses and the personal head related transfer functions in advance, and approximate $\mathbf{H_{wv}}$ from these. However, even if the individual impulse responses from the loudspeakers to the ears are exactly known, their inversion is not pratical due to the zeroes close to the unit circle. Here, the MINT concept [6] can actually solve the inversion problem with finite-length filters using several loudspeakers for a single ear, if the individual transfer functions in $\mathbf{H_{wv}}$ exhibit no common zeroes.

B. Interference cancellation. According to Eq.5, we must also cancel the interference at the ear via $\mathbf{G_{vx}}$ to satisfy

$$\mathbf{H_{wv}} * \mathbf{G_{vx}} * \mathbf{x} + \mathbf{w_n} \stackrel{!}{=} \mathbf{0}. \quad (7)$$

For that, we have to derive reference information about the undesired noise at the ear, $\mathbf{w_n}$, from $\mathbf{x}$. This requires that the noise components at the ears $\mathbf{w_n}$ must result from a coherent sound field

[1] By defining $\mathbf{y} = \mathbf{A} * \mathbf{x}$ as a matrix multiplication with elementwise convolution, the elements $y_i(k)$ of $\mathbf{y}$ are given by $y_i(k) = \sum_{j=1}^{N} \sum_{n=-\infty}^{\infty} a_{ij}(k-n)x_j(n)$ assuming that the impulse response $a_{ij}(k)$ is time-invariant. The inverse $\mathbf{A}^{-1}$ of matrix $\mathbf{A}$ is defined by $\mathbf{A}^{-1} * \mathbf{A} = \mathbf{I} \cdot \delta(k)$, with $\mathbf{I}$ as identity matrix. and $\delta(k)$ as discrete-time unit impulse. For rank-deficient or non-square matrices $\mathbf{A}$, $\mathbf{A}^{-1}$ is the pseudoinverse (see [4]).

so that we can use the matrix $\mathbf{H_{xw}}$ to describe their relation to the noise components at the microphones:

$$\mathbf{x_n} = \mathbf{H_{xw}} * \mathbf{w_n}. \tag{8}$$

Now we assume that $\mathbf{G_{vx}}$ includes a linear signal separation unit $\mathbf{G_{x_nx}}$, $\mathbf{G_{vx}} = \mathbf{G_{vx_n}}\mathbf{G_{x_nx}}$, to extract $\mathbf{x_n}$ from $\mathbf{x}$, so that

$$\mathbf{x_n} \quad = \quad \mathbf{G_{x_nx}} * \mathbf{x}. \tag{9}$$

This assures that the cancellation signal only contains noise components from $\mathbf{x}$ and that the cancellation condition can be written as

$$\mathbf{H_{wv}}\mathbf{G_{vx_n}}\mathbf{G_{x_nx}}\mathbf{H_{xw}}\mathbf{w_n} = -\mathbf{w_n}. \tag{10}$$

If this should be met independently of the actual noise signals, $\mathbf{G_{vx}}$ has to satisfy

$$\mathbf{G_{vx}} = -\mathbf{H_{wv}^{-1}}\mathbf{H_{xw}^{-1}}, \tag{11}$$

which implies always noncausality for $\mathbf{G_{vx}}$, as $\mathbf{H_{wv}}, \mathbf{H_{xw}}$ represent propagation paths of wave fields. For realizable causal $\mathbf{G_{vx}}$, some extra delay must be introduced, and therefore the cancellation can only be achieved for predictable noise signals at the ears, as becomes obvious from Eq.10. Aside from periodic signals, this is also given if the actual noise sources are close enough to the microphones: If the propagation delay from the original source to the ears is larger than the combined propagation delay from the original sources to the microphones and from the loudspeakers to the ears (plus the processing delay necessary in real DSP systems), then the predicted signals can be used as input to $\mathbf{G_{vx_n}}$ instead of $\mathbf{x_n}$. (Note that this needs to be fulfilled for all acoustic paths, and obviously, does not comply with our typical application scenarios.) Even with the causality problem solved, the determination of $\mathbf{H_{wv}}$ and the predictors still constitute blind problems as in general no reference signals at the sources for $\mathbf{n}$ and the sink $\mathbf{w}$ are available. Note also that Eq.7 describes the well-known active noise cancellation problem [7, 8], where great efforts are made to obtain useful reference signals close to the noise sources and at the sink, and success is usually limited to low frequencies and/or well-defined - mostly stationary - interference.

For the second case, where the local human sources S_i should also be cancelled, the noise term $\mathbf{w_n}$ in the above derivation must only be complemented by the additional speech term $\mathbf{w_s}$, and we have to assume that $\mathbf{G_{vx}}$ includes signal extraction units for both $\mathbf{x_n}, \mathbf{x_s}$. The fundamental problem for $\mathbf{G_{vx}}$ becomes thereby even harder, because, then, typically, human utterances such as speech signals need to be predicted.

Although hands-free, distant-listening to audio signals via loudspeakers was always a natural concept and evolved in parallel with headsets from the very beginning, according current reproduction systems do not solve the dereverberation problem nor do they cancel interference. Stereo or multichannel reproduction systems, such as 5.1 surround systems, do typically not account automatically for the specific local acoustic environment, $\mathbf{H_{wv}}, \mathbf{w_n}$. Instead, the user is allowed to choose his own frequency-dependent gain equalization, which corresponds to defining the elements of the main diagonal of $\mathbf{G_{vu}}$, and some mixing of input channels u_i, so that nondiagonal elements in $\mathbf{G_{vu}}$ are defined. With such systems, a fully controlled listening experience is therefore only possible in an anechoic, noise-free environment, and only at a certain position ('sweet spot'), which can be controlled by inserting appropriate delays into $\mathbf{G_{vu}}$.

Wave field synthesis [9, 10] overcomes the sweet spot problem by creating a well-defined sound field within an extended area - so far in two spatial dimensions - using a large number of loudspeakers (from dozens to hundreds), so that one may move freely while listening and many listeners can enjoy the same spatial realism, e.g., in a cinema. The desired transfer characteristics matrix H_{d} is here typically filled with impulse responses, so that virtual as well as remote real acoustic environments (e.g. concert halls, churches) can be reproduced. The inversion of the local acoustics $\mathbf{H_{wv}}$ can here be tackled by creating wave fields that explicitly cancel reflected waves at all points of the listening space [11].

The second problem, active interference cancellation, has been investigated for many years. However, for the given scenario, the author is not aware of any successful concepts. Again, wave field synthesis may have some potential here, as it does not rely on reference information from many points in space but reconstructs wave fields from sampled closed contours [12].

Aside from all these efforts, the apparent imperfections in reproduction technology at the acoustic human/machine interface are not prohibiting the acceptance and widespread enjoyment of present products by the human listeners. This may partly be explained by psychoacoustics, so that the listeners do not always ask for optimality subject to criteria as defined by system theory. Removal of natural-sounding reverberance and local interference in a listening space will often not even be expected. Moreover, suitable material for high quality reproduction usually undergoes careful treatment by experienced and creative sound engineers so that psychoacoustics is heavily exploited to enhance the listening experience, and realism as desired by the above signal processing criteria will not be the first priority. Therefore, it remains an open question to what extent the criteria in Eqs.6,7 must be met and whether they should be modified to better fit psychoacoustics.

2.2. Acquisition

On the audio acquisition side, the acquired desired signals must typically be suitable for reproduction in other listening spaces or for recognition or interpretation by machines, and therefore the undesired signal components in the output signal $\mathbf{z}$ are usually much less tolerable than for the listeners in $\mathbf{w}$. Although single-microphone systems were always limited in their capability to suppress unwanted noise, reverberation, and echoes without distorting the desired signals, multichannel acquisition gained momentum only in the last few years with the availability of increasing and inexpensive signal processing power and the increasing demand for hands-free interfaces for various applications.

The aim for signal acquisition is obiously to extract a vector $\mathbf{z}$ containing P separate and delayed source signals $z_i(k) \approx s_j(k) * \delta(k-k_0), (i = 1, \ldots, P; j \in \{1, \ldots, M\})$, where the delay $k_0 \geq 0$ is chosen to ensure causality of $\mathbf{G_{zx}}$. Considering Eq.3, this requires that the acoustic echoes of the loudspeaker signals must be compensated, the contributions from local noise sources and the respective other sources $s_{k\neq j}$ must be suppressed, and echoes and reverberation of the desired source s_j must be removed from the microphone signals.

For notational convenience, we assume from now on $P = M$ and disregard output permutations so that we obtain from Eq.1 as the requirement for ideal signal acquisition:

$$\mathbf{z} \quad = \quad \mathbf{G_{zu}} * \mathbf{u} + \mathbf{G_{zx}} * \mathbf{x} \stackrel{!}{=} \mathbf{s} * \delta(k - k_0).$$

To identify the individual problems, we introduce Eq.3 and note

that the decomposition of $\mathbf{x}$ presumes that we have methods to separate the components $\mathbf{H_{xs}} * \mathbf{s}, \mathbf{H_{xv}} * \mathbf{v}$, and $\mathbf{x_n}$, just as discussed for reproduction (see Eq.9):

$$\begin{aligned}\mathbf{z} &= \mathbf{G_{zu}} * \mathbf{u} + \mathbf{G_{zx}} * (\mathbf{H_{xs}} * \mathbf{s} + \mathbf{H_{xv}} * \mathbf{v} + \mathbf{x_n}) \\ &= (\mathbf{G_{zu}} + \mathbf{G_{zx}} * \mathbf{H_{xv}} * \mathbf{G_{vu}}) * \mathbf{u} + \mathbf{G_{zx}} * (\mathbf{H_{xs}} * \mathbf{s} + \mathbf{x_n}) \\ &\stackrel{!}{=} \mathbf{s} * \delta(k - k_0). \end{aligned} \quad (12)$$

From this, we can isolate three subproblems:

A. Echo cancellation. For compensating the feedback of the reproduction signals $\mathbf{u}$ into the desired signals $\mathbf{z}$, we obviously have to ask for

$$(\mathbf{G_{zu}} + \mathbf{G_{zx}} * \mathbf{H_{xv}} * \mathbf{G_{vu}}) * \mathbf{u} = \mathbf{0}. \quad (13)$$

From the viewpoint of remote communication partners sending $\mathbf{u}$ and receiving $\mathbf{z}$, this corresponds to cancelling the echoes of their own signals. For signal-independent cancellation, $\mathbf{G_{zu}}$ needs to fulfill

$$\mathbf{G_{zu}} = -\mathbf{G_{zx}} * \mathbf{H_{xv}} * \mathbf{G_{vu}}, \quad (14)$$

which is obviously a MIMO system identification problem with both input and output being observable. Note that actually only the matrix $\mathbf{H_{xv}}$ describing the acoustic paths between microphones and loudspeakers must be identified.

B. Source separation and dereverberation. In order to extract the original source signals from the convolutive mixtures in each microphone, the sources need to be separated and dereverberated such that ideally

$$\mathbf{G_{zx}} * \mathbf{H_{xs}} * \mathbf{s} = \mathbf{s} * \delta(k - k_0) \quad (15)$$

is obtained. This means that any signal-independent solution must meet:

$$\mathbf{G_{zx}} * \mathbf{H_{xs}} = \mathbf{I}_{M,M} \cdot \delta(k - k_0), \quad (16)$$

where $\mathbf{I}_{M,M}$ is the $M \times M$ identity matrix. Therefore, we have for the elements of the main diagonal of $\mathbf{G_{zx}} * \mathbf{H_{xs}}$ a multichannel blind deconvolution problem similar to Eq.6, and for the off-diagonal elements we have an interference suppression problem similar to that of Eq.17.

C. Suppression of interfering noise. To remove the local noise in the output vector $\mathbf{z}$

$$\mathbf{G_{zx}} * \mathbf{x_n} = \mathbf{0} \quad (17)$$

must be met. Signal-independent solutions would require $\mathbf{G_{zx}} = \mathbf{0}$, which, however, would also preclude the capture of the desired signals. Thus, Eq.17 actually asks only for a signal separation system fulfilling Eq.9, whose output is then subtracted from the signals $\mathbf{x}$.

Thus, similarly to signal reproduction, the subproblems in acquisition involve essentially system identification problems and signal separation/extraction problems. The separation of the various components in $\mathbf{x}$, i.e., $\mathbf{H_{xs}s}$, $\mathbf{H_{xv}v}$, and $\mathbf{x_n}$ is not only necessary to suppress the noise signals itself, it is also a necessary precondition for obtaining reference signals for the various system identification problems, i.e., for determining $\mathbf{G_{zx}}$, $\mathbf{G_{zu}}$ and $\mathbf{G_{xv}}$.

With the given spatial diversity by several microphones, the separation of $\mathbf{x}$ into its components can exploit orthogonality in both the time/frequency and the spatial domain. The spatial domain is especially important, as the involved signals are usually not orthogonal in time/frequency. This constitutes a major advantage of the multichannel approaches over single-channel approaches for acoustic human/machine interfaces. One should note, however, that in time/frequency and also in the spatial domain, apertures are finite and imply finite resolution, and sampling frequencies of the apertures are limited, which implies aliasing. For determining the optimum, usually time-varying, spatiotemporal filters for signal separation, we typically use a-priori knowledge (e.g., about source positions), heuristic detection algorithms (e.g., for speech activity at a certain time from a certain direction) or parameter estimation concepts based on (mostly short-time) signal statistics.

2.3. Localisation

The task of localizing and tracking active desired sources S_i is different from signal acquisition and reproduction, insofar, as the output is not a desired signal resulting from modifying input signals, but position information as derived from analyzing the input signals $\mathbf{x}$. Clearly, extraction of the desired signals $\mathbf{s}$ should be beneficial, and it seems obvious that knowledge of $\mathbf{H_{xs}}$ should facilitate localization. The according techniques for our scenario will be briefly reviewed when discussing recent advances for signal acquisition below.

3. SOME TECHNIQUES FOR SIGNAL ACQUISITION

Rather than attempting a comprehensive overview over the numerous solutions for the four classes of problems in signal acquisition, we present here a brief synopsis of basic techniques and some recent results biased towards the work of the author's research group.

3.1. Acoustic echo cancellation

For a convenient illustration of the basic mechanisms we assume that the sound reproduction system $\mathbf{G_{vu}}$ is transparent, $\mathbf{G_{vu}} = \mathbf{I}_{K,K} \cdot \delta(k)$, and consider the system identification problem only for a single microphone signal and a single output signal ($N = P = 1$) with $\mathbf{G_{zx}} = \delta(k)$. (The application to microphone arrays with $N > 1$ has been discussed in [13], and has been generalized in [14].) Thereby, Eq.14 reduces to $\mathbf{G_{zu}} = -\mathbf{H_{xv}}$, where the matrices are row vectors with K generally time-variant impulse responses as elements:

$$\mathbf{G_{zu}} = (g_1(k), \ldots, g_K(k)), \quad (18)$$

$$\mathbf{H_{xv}} = (h_1(k), \ldots, h_K(k)). \quad (19)$$

Using an FIR model of length L_g we obtain for the estimate of the echo (see Fig.2)

$$\widehat{y}(k) = \mathbf{g}^T(k)\mathbf{u}(k), \quad (20)$$

where

$$\mathbf{g}(k) = \left(\mathbf{g}_1^T(k), \ldots, \mathbf{g}_K^T(k)\right)^T, \quad (21)$$

$$\mathbf{u}(k) = \left(\mathbf{u}_1^T(k), \ldots, \mathbf{u}_K^T(k)\right)^T, \quad (22)$$

with the individual impulse responses and data vectors

$$\mathbf{g}_i(k) = \left(g_{i,0}(k), \ldots, g_{i,L_g-1}(k)\right)^T, \quad (23)$$

$$\mathbf{u}_i(k) = (u_i(k), \ldots, u_i(k - L_g + 1))^T, \quad (24)$$

respectively. The estimation error reads:

$$e(k) = y(k) - \widehat{y}(k), \quad (25)$$

where

$$e(k) =: z(k)|_{\mathbf{s}=\mathbf{0},x_n=0}\,,\quad y(k) =: x(k)|_{\mathbf{s}=\mathbf{0},x_n=0}\,. \tag{26}$$

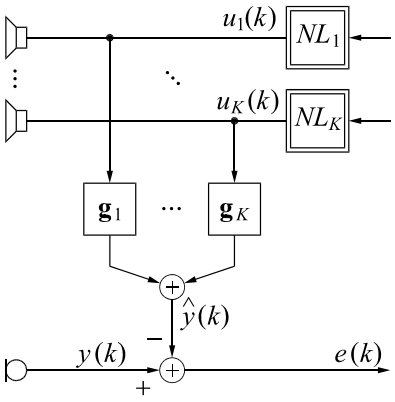

Figure 2: Echo cancellation for K-channel reproduction

In order to follow the time-variance of the impulse response $h_i(k)$, gradient-type adaptive algorithms are common to approximate the optimum Wiener solution $\mathbf{g}(k)$:

$$\mathbf{g}(k) = \mathbf{g}(k-1) + \mathbf{k}(k)e(k), \tag{27}$$

where the 'Kalman gain' vector $\mathbf{k}(k)$ determines the direction of the adaptation. While for single-channel echo cancellation ($K = 1$) simple adaptation algorithms, such as the normalized least mean square (NLMS) algorithm (corresponding to $\mathbf{k}(k) = \alpha\mathbf{u}/(\mathbf{u}^H\mathbf{u})$, $0 < \alpha < 2$, see [16]) are very popular, for multichannel echo cancellation ($K \geq 2$), algorithms with improved convergence properties are necessary. This is due to the often strong, time-varying correlation between the K input channels $u_i(k)$, which results from the fact that the signals $u_i(k)$ are usually just different mixtures of a common set of sources. As an alternative to the NLMS algorithm, the RLS ('recursive least squares') algorithm using the Kalman gain vector $\mathbf{k}(k) = \mathbf{R}_{\mathbf{uu}}^{-1}(k)\cdot\mathbf{u}$ (with $\mathbf{R}_{\mathbf{uu}}$ being the estimated autocorrelation matrix of $\mathbf{u}$) promises fastest convergence. However, even here we have to improve the condition number of $\mathbf{R}_{\mathbf{uu}}$, e.g., by an (ideally imperceptible) nonlinearity NL_i (cf. Fig.2) [17]. As alternatives to nonlinearities, time-varying allpass filters [18] and the insertion of additive noise (e.g. by an audio codec, such as MP3 or AAC, [19]) have been suggested. Obviously, none of these methods will comply with the quest for perfect reproduction, and will be especially objectionable for large numbers of channels L, where they need to be applied more rigorously in order to obtain sufficiently fast convergence.

As a direct inversion of the $K\cdot L_G \times K\cdot L_G$ matrix $\mathbf{R}_{\mathbf{uu}}^{-1}(k)$ is still unrealistic for real-time implementations with $K\cdot L_G = 1000\ldots20000$, approximative solutions in the DFT domain are very attractive. In [15] an algorithm was presented which requires only the inversion of L_G matrices of size $K\times K$ instead of one matrix of size $(K\cdot L_G)\times(K\cdot L_G)$, and thereby allows real-time operation of a $K = 5$-channel echo canceller with $K\cdot L_G > 20000$ filter coefficients on an ordinary PC (Intel 1.7GHz, dual processor board, sampling frequency 12kHz). In Fig.3 typical convergence curves of the system error norm ($\propto \log_{10}(\|\mathbf{G}_{\mathbf{zu}} + \mathbf{H}_{\mathbf{xv}}\|^2/\|\mathbf{H}_{\mathbf{xv}}\|^2)$) and the echo suppression (ERLE) are depicted for various K. The ERLE curves demonstrate that with proper parametrization echo suppression need not deteriorate with increasing channel number K. While this configuration is suitable, e.g. for voice-controlled home-theatres, further research is needed for more demanding environments, such as telecollaboration of musicians playing in distant studios, where $L > 5$ and $f_\mathrm{s} \geq 32\mathrm{kHz}$ must be expected and, additionally, the signal delay must be minimum.

As one option for a very large number of channels such as in wave field synthesis, a new echo cancellation concept based on *wave domain adaptive filtering (WDAF)* [11, 20] is able to perform echo cancellation in a transform domain with eigenfunctions of the sound fields as basis functions. Aside from many loudspeakers, this requires also a large number of microphones (e.g., $L = N = 48$ in an experimental systems) to sample the wave fields and provide the reference information for adapting the echo path models [20].

In some common applications, especially with low-cost loudspeakers and low-power amplifiers, the linear model for the feedback path $\mathbf{H}_{\mathbf{vx}}$ is not valid any more. In [21], the matrix notation as used so far for linear systems was extended to incorporate Volterra filters, and an efficient DFT domain algorithm was presented which allows modelling of loudspeaker nonlinearities by second-order Volterra filters [22]. Although nonlinearities in loudspeakers and amplifiers are not really memoryless, it was shown, that echo path models with a memoryless nonlinearity can still be effective in some practical cases [23]. Especially so-called *power filters*, where the signal samples are raised to different powers and the resulting sequences are then passed through parallel linear filters, can provide effective models for nonlinearities as they may occur in our scenario [24].

3.2. Signal extraction and interference suppression

In the following we consider multichannel techniques for determining spatiotemporal filters $\mathbf{G}_{\mathbf{zx}}$ to approximate Eqs.17, 15 and/or 16. Seemingly unrelated at first glance, they pursue the same goals and differ essentially regarding the used reference information and optimization criteria.

3.2.1. Beamforming

Beamforming microphone arrays aim at both the signal separation and the suppression of noise and interference, and ideally extract undistorted desired source signals. A general treatment of theoretical concepts, alternative approaches, and other aspects of design and applications can be found, e.g., in [25, 26, 27]. Beamforming essentially forms a 'beam' of increased sensitivity towards the location of a desired source and simultaneously tries to suppress all other sources. If not known, the position of the desired source must be determined by localization methods as discussed below. If several desired sources need to be extracted, several beamformers can work in parallel using the same microphone signals [28], however, for adaptive beamformers involving estimation of statistical quantities for individual sources the estimation will suffer from the interference of additional sources.

For a single desired source, the components x_i of $\mathbf{x}$ are in the simplest case individually delayed and summed such that components of the desired source signal are summed up coherently while signals from other locations are summed with generally nonzero phase differences and cancel out to a certain degree ('delay and sum beamformer', DSB). This supposes that the location or at least

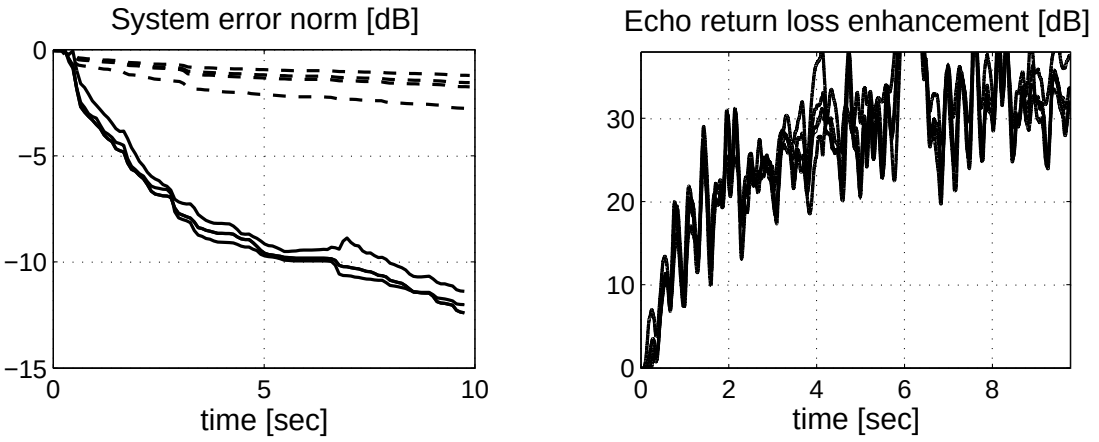

Figure 3: Convergence of DFT-domain adaptation after [15] for music signal reproduction with $K = 2, 3, 4, 5$ channels. System error norm (left) relative to NLMS(dashed lines), echo return loss enhancement (ERLE, right)

the direction of arrival (DOA) of the desired source is known. Using filters instead of delays, 'filter and sum' beamformers are obtained, which allow a frequency-selective modification of the spatial filtering characteristics of the plain DSB. Such beamformers are still data-independent as long as they do not account for the actual signal statistics of $\mathbf{x}$.

Current techniques for our scenario use data-dependent beamformers, where the spatiotemporal filtering $\mathbf{G_{zx}}$ is mostly designed to either pursue a minimum mean square error (MMSE) criterion or to aim at minimum output variance while ensuring distortionless response (MVDR) for the desired source [29][2]. MMSE criteria lead to a multichannel Wiener filter solution, with the inherent problem that the desired signal will be distorted while the suppression of noise and interference is maximized [30].

On the other hand, MVDR criteria ensure an undistorted desired signal as long as the source position is exactly known, by imposing a constraint on the optimization. An example for an array response, i.e., the magnitude gain as a function of DOA and frequency, is shown in Fig.4. As can be seen, the interferer is strongly suppressed, while the desired direction $\Theta = 0$ is not attenuated. Note also that at low frequencies there is almost no attenuation of signals from any direction, which is due to the limited size of the aperture of the microphone array relative to the wavelength. To circumvent the constraint optimization problem, the so-called Generalized Sidelobe Canceller (GSC) has been proposed [32], which will cancel the desired signal, however, if the target direction is not precisely known. A robust version of the GSC as depicted in Fig.5 has been developed in [33] and further been refined using a DFT-domain implementation [34]. The GSC principle [32] foresees that a signal-independent beamformer $\mathbf{c}$ filters the sensor signals so that the desired signal arriving via the direct path remains undistorted, whereas, ideally, other directions should be suppressed. In the lower path, an adaptive blocking matrix $\mathbf{B}$ aims at suppressing all components originating from the desired signal s_i, so that only noise components appear at the output of $\mathbf{B}$, and thereby essentially approximates Eq.9. From the outputs of $\mathbf{B}$ the adaptive interference canceller $\mathbf{a}$ derives an estimate for

[2]Note that the MMSE and MVDR criterion is defined for stationary processes and based on statistical averaging whereas for nonstationary processes and real data samples, the criteria must be modified, thereby offering many variations [26]

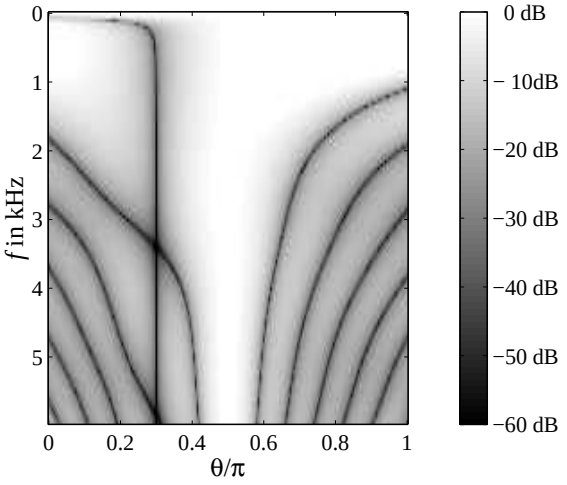

Figure 4: *Array response (*beam pattern*) of an MVDR beamformer for a linear equispaced array with $M = 9$ sensors, with spacing $d = 4$cm, to an interferer emitting white noise from $\Theta = 0.3\pi$ (with permission from [26]).*

the remaining noise component in the output of $\mathbf{c}$, by minimizing an estimate of the variance of the output z_i. Obviously, the fixed beamformer $\mathbf{c}$ and the interference canceller $\mathbf{a}$ jointly perform interference suppression in the sense of Eq.17. The resulting signal z_i will also be slightly dereverberated relative to $\mathbf{G_{xs}} * \mathbf{s}$ as the fixed beamformer $\mathbf{b}$ will attenuate reflections arriving from attenuated angles of incidence.

As for the separation of the noise components, a time-variant blocking matrix $\mathbf{B}$ can use spatial, spectral, and temporal selectivity to isolate the noise and suppress the desired signal. The adaptation of the blocking matrix $\mathbf{B}$ allows to follow movements of the desired source S_i and thereby provides robustness against desired signal cancellation: Otherwise, if the desired signal leaks through the blocking matrix, it will be treated as a noise component and subtracted from the output of $\mathbf{c}$. The spatial selectivity of $\mathbf{B}$ is very beneficial as it allows to completely suppress the signal arriving from the assumed source direction, but it usually cannot

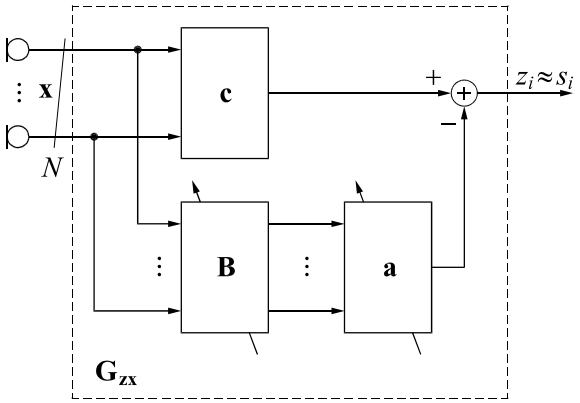

Figure 5: *Structure of a robust Generalized Sidelobe Canceller.*

completely suppress reverberation of the desired signal. Therefore, adaptation of the blocking matrix $\mathbf{B}$ has to exploit temporal selectivity: It should only be adapted during periods when the desired signal is dominant. Likewise, the interference canceller $\mathbf{a}$ should only be adapted when noise and interference are dominant.

While the original proposal [33] suggests an implementation by FIR filters in the time domain, both blocking matrix and interference canceller become significantly more efficient and robust if spatial selectivity and the temporally selective adaptation is combined with spectral selectivity: Realizing the entire structure in the DFT domain allows bin-selective decisions and filter adaptation and improves performance significantly, especially for nonstationary noise and interferers [34, 35]. For a linear array of $N = 8$ sensors with spacing 4cm, more than 20dB of interference suppression with negligible distortion of the desired signal can be obtained in environments with moderate reverberation ($T_{60} = 0.3$sec).

It should be mentioned that for applications where only few microphones can be used and the aperture must be very small, so-called differential arrays [36] are a natural choice. Their properties can be derived as special cases of the general beamforming concepts.

Major application areas where microphone arrays recently gained popularity include hands-free car telephony, video-conferencing equipment for both desktops and studios, and front-ends for hands-free automatic speech recognition. For future applications of wave field synthesis, it might be of interest that, similarly as for multichannel echo cancellation, the interference cancellation concept of adaptive beamformers can be carried over to the wave domain [37].

3.2.2. Blind source separation and blind deconvolution

When the desired source position is not available and the signal extraction should not rely on a well-defined array geometry as with beamforming, blind signal processing algorithms are especially attractive. Unlike the original blind source separation (BSS) concepts, which separated scalar signal mixtures [38], in our scenario, BSS algorithms have to separate convolutive mixtures given by $\mathbf{H_{xs}} * \mathbf{s}$, so that the output signals are usually still linearly filtered versions of the original signals. On the other hand, dereverberation by blind deconvolution aims at extracting the original desired signals by additionally assuming a source model for the desired signals. BSS can be understood as blind beamforming [39], and blind deconvolution algorithms would then correspond to blind beamforming with additional equalization of the acoustic channel from the source to the microphones.

Separating convolutive mixtures of several desired sources, means that BSS aims at $\mathbf{G_{zx}} * \mathbf{H_{xs}} * \mathbf{s} = \mathbf{z} \approx \mathbf{s}$. Here, the $\approx$ sign does allow for an additional filtering of each vector element but not for mixing of the vector elements. The problem is illustrated in Fig.6 for $M = N = 2$. From that it can be seen, that BSS realizes a GSC-like interference cancellation for each output z_i [40], however, due to the blindness, $\mathbf{G_{zx}}$ cannot be determined by the same criterion. Lacking reference information, BSS essentially attempts to minimize statistical dependency ('minimum mutual information') between the output signals z_i, but it should be emphasized that the separation performance of the resulting filters in $\mathbf{G_{zx}}$ is nevertheless determined by the spatial selectivity of $\mathbf{G_{zx}}$. Note that the optimization criteria of BSS do not address

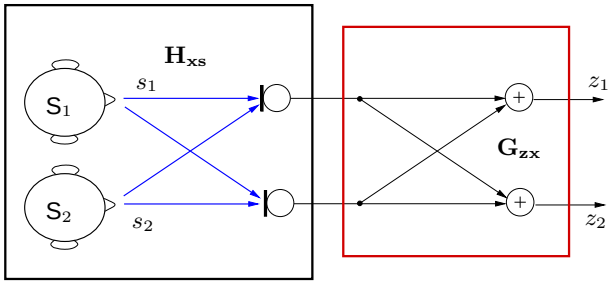

Figure 6: *Signal model for BSS with $M = N = 2$.*

the dereverberation problem Eq.16, although the spatial selectivity of the resulting $\mathbf{G_{zx}}$ may contribute to dereverberation (just as beamforming does).

For the given convolutive mixtures of speech and audio signals, three stochastic signal properties can be exploited to determine optimum demixing filters $\mathbf{G_{zx}}$ [41, 42, 43]:

Nonwhiteness of speech and audio signals can be exploited by simultaneous block-diagonalization of correlation matrices formed by $z_i(k), z_j(k-d)$, for all relative delays d.

Nonstationarity can be exploited by simultaneous diagonalization of several short-time estimates of the correlation matrices, assuming that the optimum filters vary less than the short-time signal statistics.

Nongaussianity can exploited by higher order statistics (HOS) as used for independent component analysis (see, e.g., [44]). Then, instead of minimizing crosscorrelation matrices between different channels, joint probability density functions linking the samples of different channels $z_i(k), z_j(k-d)$ must be factorized across different channels while leaving the joint pdfs of the samples within a channel unchanged.

For most known algorithms, only one or two of these properties are exploited. Successful systems have been presented that are based on second order statistics (SOS) only, and use nonwhiteness and nonstationarity only [45, 46, 43]. Recently, TRINICON has been presented as a generic algorithm, which simultaneously exploits all three properties and minimizes mutual information [41, 42, 47]. Here, spherical invariant random processes (SIRPs) [48] can be incorporated into the score function to provide an efficient model for multivariate pdfs of speech signals.

As in our scenario convolutive mixtures have to be separated, an implementation in the DFT domain is especially attractive, because it converts convolutive mixtures in the time domain into scalar mixtures for each frequency bin. However, if separation in frequency bins is carried out independently, this leads to the so-called internal permutation problem: the separated DFT bins for sources S_i and S_j cannot be aligned to guarantee that all bins with components of a source S_i appear at the same output of the BSS system. Moreover, most frequency-domain algorithms are implicitly based on the DFT-inherent circular convolution of the input data instead of the required linear convolution. Heuristic repair mechanisms are common and sometimes reasonably efficient [45, 49]. On the other hand, within the framework of a generic SOS or HOS algorithm, time-domain criteria can also be transformed rigorously into the DFT domain and, thereby, both problems are solved perfectly [41].

In Fig.7 the convergence of the signal-to-interference power ratio for various off-line BSS algorithms for $M = N = 2$ and demixing filters of length 512 is compared (for details see [42]). The speech signal mixtures were recorded in a real room with $T_{60} = 0.15$sec at a sampling frequency of 16kHz. Obviously,

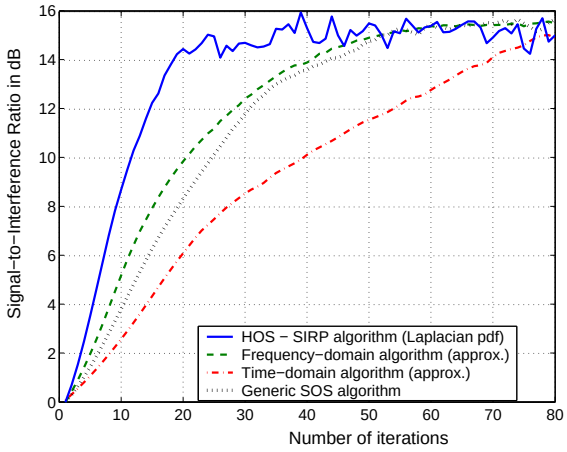

Figure 7: *Convergence curves for off-line BSS (from [42]).*

the HOS-SIRP algorithm [47], which accounts for all three signal properties, clearly outperforms the other algorithms. The generic SOS exhibits roughly the same convergence speed as the well-known frequency domain algorithm [45], which is based on heuristic repair mechanisms for the internal permutation and the circular convolution problem and turns out to be an approximation of the generic SOS algorithm. The relation of the time-domain approximation to the generic SOS algorithm corresponds to the relation of the NLMS to the RLS adaptation algorithm, which explains the somewhat slower convergence. However, this approximation permitted the first known real-time implementation of a time-domain algorithm which perfectly avoids internal permutation and circular convolution [50], whereas previously reported real-time implementations of BSS systems all operate in the DFT domain (e.g., [45, 51]).

Research in BSS strives and, recently, a BSS system for up to $P = 6$ channels has been demonstrated in real-life situations [52]. Moreover, noise-robust versions have also been published already [53]. One of the major challenges will be the handling of $M > N$, i.e., if more sources than available microphones have to be separated.

The generic TRINICON concept provides also a promising means to tackle dereverberation [42]. Fig.8 illustrates that the problem actually only asks for (multichannel) partial blind deconvolution (MCPBD) where the filtering by the human vocal tract has to be preserved, as otherwise, the output would be the signal as produced by the glottis. In Fig.9 an example for the dereverberation capability of the TRINICON-based SOS-MCBPD algorithm relative to SOS-BSS and DSB is shown. Although this result

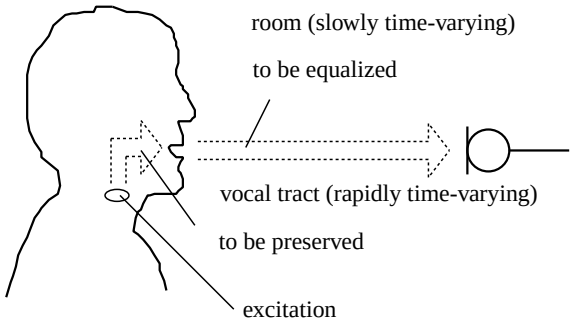

Figure 8: *Dereverberation and the MCPBD principle.*

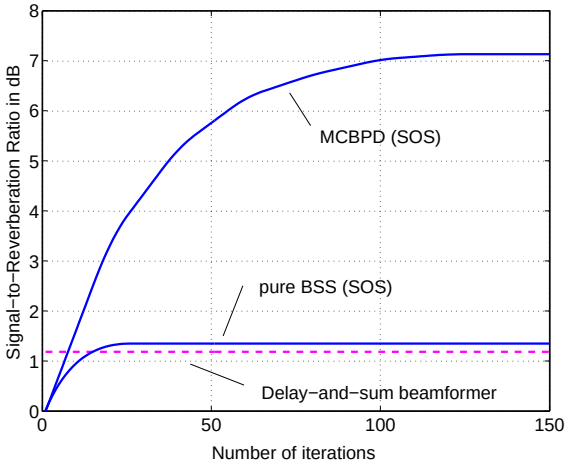

Figure 9: Dereverberation by 2×2 SOS - MCBPD, $L = 1024$, $T_{60} \approx 200$ms, $f_S = 16kHz$

shows the potential of this kind of algorithms, a robust solution for real-time dereverberation as, e.g., desired for distant-talking speech recognition applications presents still a major challenge.

3.2.3. Localization

Traditional methods for source localization of sound sources in reverberant rooms follow either one or a combination of the following concepts: a, Steered response beamforming, b, TDOA estimation by crosscorrelation measurement, or c, spectral analysis from array processing techniques [54]. Steered response beamforming essentially scans the acoustic space for peaks of signal power to locate sources. This involves relatively high computational load if localization should be precise. Moreover, it may

easily misinterpret focal points of reflections and noise as desired sources. Crosscorrelation-based methods detect peaks in the generalized cross-correlation (GCC) of microphone pairs and compute from the corresponding time differences of arrival (TDOA) the source locations [55]. As it is computationally relatively inexpensive, it is very popular and performs well for low noise and low-reverberation environments as long as only a single source must be detected. However, room reverberation and noise can only be accounted for by tuning window lengths and weighting mechanisms. The main idea of statistical array processing is to decompose the correlation matrix of the sensor signals into its eigenvectors and to use the M eigenvectors corresponding to the largest eigenvalues as indicators for the desired source locations. Based on this subspace idea, wide classes of algorithms have been derived (e.g. MUSIC, ROOT-MUSIC, ESPRIT) [56], which are inherently based on a narrowband signal model and rely strongly on well-established correlation matrices, which in turn require sufficiently stationary environments (as they are rarely given in our scenario).

Recently, new concepts have been proposed that explicitly address wideband sources and nonstationary acoustic environments. Most notably, the adaptive eigenvalue decomposition [57] uses a microphone pair to approximately identify the acoustic paths to a source. From the resulting impulse responses only the dominant peaks are considered to obtain a useful TDOA estimate. As opposed to GCC, this method thus explicitly accounts for the reverberance of the room. From the BSS concept of Fig.6, we see that detecting the dominant peaks in the impulse responses of the demixing filters in $\mathbf{G}_{zx}$ yields equivalent TDOA estimates even for two sources [58].

Aiming at even larger number of wideband sources, the above array processing methods have recently been applied to signals transformed to the wave domain [59, 60], where the array processing algorithms behave just as for narrowband signals.

Beyond estimating instantaneous source locations, tracking of moving sources can be supported by movement models, such as extended Kalman filters [62] or particle filters [61].

4. SUMMARY AND CONCLUSIONS

For our discussion of the various signal processing problems at a generic acoustic human/machine interface we distinguished signal reproduction and signal acquisition and first stated the fundamental problems before discussing recent solutions. On the reproduction side, wave field synthesis overcomes the sweet-spot problem of traditional reproduction systems and offers some potential to solve other fundamental reproduction problems: Both listening room equalization, and noise and interference compensation are under investigation, but wide-band wide-range active noise compensation appears to be out of reach. For signal acquisition, acoustic echo cancellation as a non-blind MIMO system identification problem, appears close to being solved, although for multichannel reproduction system still fundamental problems await elegant solutions. Over the last few years, adaptive beamforming has reached a certain maturity in achieving the desired signal separation and interference cancellation. Without relying on source location information and due to a more powerful optimization criterion, blind source separation techniques offer significant potential for the same tasks as traditional beamforming. Dereverberation, involving blind deconvolution, remains a major challenge for the coming years especially with regard to its long-desired application to distant-talking speech recognition. Finally, new promising localization techniques for nonstationary wideband sources appear as by-products of blind signal separation and wave-domain signal representation, and illustrate the close relations between the different problems and solutions.

In summary, we may safely conclude that despite of significant progress over the last few years, many fascinating challenges for digital signal processing on both theoretical and experimental level remain on the way to an ideal generic acoustic human/machine interface.

5. ACKNOWLEDGEMENTS

This paper would not have been written without the underlying research contributions of the author's PhD students, notably, Robert Aichner, Herbert Buchner, Wolfgang Herbordt, Fabian Küch, and Heinz Teutsch.

6. REFERENCES

[1] J.L. Flanagan, J.D. Johnston, R. Zahn, and G.W. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," *J. Acoust. Soc. Am.*, vol. 78, no. 5, pp. 1508–1518, Nov. 1985.

[2] J.L. Flanagan, D.A. Berkley, G.W. Elko, J.E. West, and M.M. Sondhi, "Autodirective microphone systems," *Acustica*, vol. 73, pp. 58–71, 1991.

[3] W. Kellermann, H. Buchner, W. Herbordt, and R. Aichner, "Multichannel acoustic signal processing for human/machine interfaces - fundamental problems and recent advances," in *Proc. Int. Conf. on Acoustics (ICA)*, Kyoto, Apr. 2004.

[4] C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs, NJ, 1974.

[5] H. Kuttruff, *Room Acoustics*, Spon Press, London, 4th edition, 2000.

[6] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 36, no. 2, pp. 145–152, Feb. 1988.

[7] P.A. Nelson and S.J. Elliott, *Active Control of Sound*, Academic Press, London, 1992.

[8] S. M. Kuo and D. R. Morgan, *Active Noise Control Systems*, Wiley, New York, 1996.

[9] P.J. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wavefield synthesis," *J. Acoust. Soc. Am.*, vol. 93, no. 5, pp. 2764–2778, May 1993.

[10] S. Spors, H. Teutsch, A. Kuntz, and R. Rabenstein, "Sound field synthesis," in *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, J. Benesty and Y. Huang, Eds., pp. 323–344. Kluwer Academic Publishers, Boston, Feb. 2004.

[11] S. Spors, H. Buchner, and R. Rabenstein, "A novel approach to active listening room compensation for wave field synthesis using wave-domain adaptive filtering," in *Proceedings Intern. Conference on Acoustics, Speech, and Signal Processing*, Montreal, Canada, May 2004, IEEE.

[12] A. Kuntz and R. Rabenstein, "An approach to global noise control by wave field synthesis," in *Signal Processing XII: Theories and Applications (Proceedings of EUSIPCO-2004)*, Vienna, Austria, Sept. 2004, EURASIP.

[13] W. 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, Berlin, May 2001.

[14] H. Buchner, W. Herbordt, and W. Kellermann, "An efficient combination of multi-channel acoustic echo cancellation with a beamforming microphone array," in *Proceedings of the Workshop on Hands-free Speech Communication*, Kyoto, Japan, Apr. 2001, IEEE, pp. 55–58.

[15] H. Buchner, J. Benesty, and W. Kellermann, "Multichannel frequency-domain adaptive filtering with application to acoustic echo cancellation," in *Adaptive Signal Processing: Application to Real-World Problems*, J. Benesty and Y. Huang, Eds., chapter 3. Springer, Berlin, Jan. 2003.

[16] C. Breining, P. Dreiseitel, E. Hänsler, A. Mader, B. Nitsch, H. Puder, T. Schertler, G. Schmidt, and J. Tilp, "Acoustic echo control," *IEEE Signal Processing Magazine*, vol. 16, no. 4, pp. 42–69, July 1999.

[17] 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. on Speech and Audio Processing*, vol. 6, no. 2, pp. 156–165, Mar. 1998.

[18] K. Sugiyama, Y. Joncours, and A. Hirano, "A stereo echo canceller with correct echo-path identification on an input sliding technique," *IEEE Trans. on Signal Processing*, vol. 49, no. 11, Nov. 2001.

[19] T. Gänsler and P. Eneroth, "Influence of audio coding on stereophonic acoustic echo cancellation," in *Proceedings Intern. Conference on Acoustics, Speech, and Signal Processing*, Seattle, WA, May 1998, IEEE, pp. 3649–3652.

[20] H. Buchner, S. Spors, and W. Kellermann, "Wave-domain adaptive filtering: Acoustic echo cancellation for full-duplex systems based on wave-field synthesis," in *Proceedings Intern. Conference on Acoustics, Speech, and Signal Processing*, Montreal, Canada, May 2004, IEEE.

[21] F. Kuech, W. Kellermann, H. Buchner, and W. Herbordt, "Acoustic signal processing for distant-talking speech recognition: Nonlinear echo cancellation in a generic multichannel interface," in *Proc. Workshop on Nonlinear Signal and Image Processing (NSIP'03)*, Grado, Italy, June 2003, IEEE-EURASIP.

[22] F. Kuech and W. Kellermann, "Partitioned block frequency-domain adaptive second-order volterra filter," *IEEE Trans. on Signal Processing*, vol. 53, no. 2, pp. 564–575, Feb. 2005.

[23] A. Stenger and W. Kellermann, "Adaptation of a memoryless preprocessor for nonlinear acoustic echo cancelling," *Signal Processing*, vol. 80, pp. 1747–1760, sep 2000.

[24] F. Kuech, A. Mitnacht, and W. Kellermann, "Nonlinear acoustic echo cancellation using adaptive orthogonalized power filters," in *Proceedings Intern. Conference on Acoustics, Speech, and Signal Processing*, Philadelphia, USA, March 2005, IEEE.

[25] M.S. Brandstein and D. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*, Springer, Berlin, 2001.

[26] W. Herbordt and W. Kellermann, "Adaptive beamforming for audio signal acquisition," in *Adaptive Signal Processing: Application to Real-World Problems*, J. Benesty and Y. Huang, Eds., chapter 6, pp. 155–194. Springer, Berlin, Jan. 2003.

[27] H.L. van Trees, *Optimum Array Processing*, Wiley, New Yrok, NY, 2002.

[28] W. Kellermann, "A self-steering digital microphone array," in *Proceedings Intern. Conference on Acoustics, Speech, and Signal Processing*, Toronto, Ontario, Canada, May 1991, IEEE, pp. 3581–3584.

[29] B.D. Van Veen and K.M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP magazine*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[30] 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 13, pp. 19–38. Springer, Berlin, May 2001.

[31] S.Doclo and M. Moonen, "GSVD-based optimal filtering for multi-microphone speech enhancement," in *Microphone Arrays: Signal Processing Techniques and Applications*, M.S. Brandstein and D. Ward, Eds., chapter 6, pp. 111–132. Springer, Berlin, May 2001.

[32] L.J. Griffiths and C.W. Jim, "An alternative approach to linear constrained adaptive beamforming," *IEEE Transactions on Antennas and Propagation*, vol. 30, no. 1, pp. 27–34, Jan. 1982.

[33] O. Hoshuyama and A. Sugiyama, "A robust adaptive beamformer with a blocking matrix using constrained adaptive filters," in *Proceedings Intern. Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1996, pp. 925–928.

[34] 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. 21–31, Jan. 2003.

[35] W. Herbordt, T. Trini, and W. Kellermann, "Robust spatial estimation of the signal-to-interference ratio for non-stationary mixtures," in *Conf. Rec. of the Seventh International Workshop on Acoustic Echo and Noise Control (IWAENC 03)*, Kyoto, Sept. 2003.

[36] G. Elko, "Differential microphone arrays," in *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, Y. Huang and J. Benesty, Eds., pp. 2–65. Kluwer, 2004.

[37] W. Herbordt, S. Nakamura, S. Spors, H. Buchner, and W. Kellermann, "Wave field cancellation using wave-domain adaptive filtering," in *Conf. Rec. Joint Workshop for Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, Piscataway, USA, March 2005.

[38] A.J. Bell and T.J. Sejnowski, "An information-maximisation approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 10004–1034, July 1995.

[39] J.-F. Cardoso and A. Souloumiac, "Blind beamforming for non-gaussian signals," *IEE Proceedings-F*, vol. 140, no. 6, pp. 362–370, Dec. 1993.

[40] S. Araki, S. Makino, Y. Hinamoto, R. Mukai, T. Nishikawa, and H. Saruwatari, "Equivalence between frequency-domain blind source separation and frequency-domain adaptive beamforming for convolutive mixtures," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 11, pp. 1157–1166, Oct. 2003.

[41] H. Buchner, R. Aichner, and W. Kellermann, "Blind source separation for convolutive mixtures: A unified treatment," in *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, J. Benesty and Y. Huang, Eds. Kluwer Academic Publishers, Boston, Feb. 2004.

[42] H. Buchner, R. Aichner, and W. Kellermann, "TRINICON: A versatile framework for multichannel blind signal processing," *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, May 2004, accepted.

[43] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 120–134, Jan. 2005.

[44] J.-F.Cardoso, "Blind signal separation: Statistical principles," *Proceedings of the IEEE*, vol. 86, no. 10, pp. 2009–2025, Oct. 1998.

[45] L. Parra and C. Fancourt, "An adaptive beamforming perspective on convolutive blind source separation," in *Noise Reduction in Speech Applications*, G. Davis, Ed. CRC Press LLC, 2002.

[46] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of a class of blind source separation algorithms blind source separation for convolutive mixtures," in *Proc. Int. Symp. on Independent Component Analysis (ICA)*, Nara, Japan, Apr. 2003.

[47] H. Buchner, R. Aichner, and W. Kellermann, "Blind source separation for convolutive mixtures exploiting nongaussianity, nonwhiteness, and nonstationarity," in *Conf. Rec. of the Seventh International Workshop on Acoustic Echo and Noise Control (IWAENC 03)*, Kyoto, Sept. 2003.

[48] H. Brehm and W. Stammler, "Description and generation of spherically invariant speech-model signals," *Signal Processing*, vol. 12, pp. 119–141, 1987.

[49] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 530–538, Sept. 2005.

[50] R. Aichner, H. Buchner, Fei Yan, and W. Kellermann, "Real-time convolutive blind source separation based on broadband approach," *Proc. Int. Symposium on Independent Component Analysis and Blind Signal Separation (ICA)*, Sept. 2004.

[51] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Real-time blind source separation for moving speakers using blockwise ica and residual crosstalk-subtraction," in *Proc. Int. Symp. on Independent Component Analysis (ICA)*, Nara, Japan, Apr. 2003.

[52] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Blind source estimation and DOA estimation using small 3-D microphone array," in *Conf. Rec. Joint Workshop for Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, Piscataway, USA, March 2005.

[53] R. Aichner, H. Buchner, and W. Kellermann, "Convolutive blind source separation for noisy mixtures," Strasbourg, France, March 2004.

[54] J.H. DiBiase, H.F. Silverman, and M.S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays: Signal Processing Techniques and Applications*, M.S. Brandstein and D. Ward, Eds., chapter 8, pp. 157–180. Springer, Berlin, May 2001.

[55] C.H. Knapp and G.C. Carter, "The generalized correlation method for estimation of time-delay," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 24, pp. 320–327, Aug. 1976.

[56] H. Krim and M. Viberg, "Two decades of array signl processing research - the parametric approach," *IEEE Signal Processing Magazine*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[57] J. Benesty, "Adaptive eigenvalue decomposition for passive source localization," *J. Acoust. Soc. Am.*, vol. 107, no. 1, pp. 384–391, Jan. 2000.

[58] H. Buchner, R. Aichner, J. Stenglein, H. Teutsch, and W. Kellermann, "Simultaneous localization of multiple sound sources using blind adaptive mimo filtering," in *Proceedings Intern. Conference on Acoustics, Speech, and Signal Processing*, Philadelphia, USA, March 2005, IEEE.

[59] H. Teutsch and W. Kellermann, "EB-ESPRIT: 2D Localization of multiple wideband audio sources using Eigen-Beams," in *Proceedings Intern. Conference on Acoustics, Speech, and Signal Processing*, Philadelphia, USA, March 2005, IEEE.

[60] H. Teutsch and W. Kellermann, "Eigen-Beam Processing for Direction-of-Arrival Estimation Using Spherical Apertures," in *Conf. Rec. Joint Workshop for Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, Piscataway, USA, March 2005.

[61] D. Ward, E. Lehmann, and R. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 11, pp. 826–836, Nov. 2003.

[62] N. Strobel, S. Spors and R. Rabenstein, "Joint Audio-Video Signal Processing for Object Localization and Tracking," in *Microphone Arrays: Signal Processing Techniques and Applications*, M.S. Brandstein and D. Ward, Eds., pp. 203–225. Springer, Berlin, May 2001.

[63] H. Buchner, S. Spors, W. Kellermann, and R. Rabenstein, "Full-duplex communication systems with loudspeaker arrays and microphone arrays," in *Proc. Int. Conf. on Multimedia and Expo (ICME)*, Lausanne, Switzerland, Aug. 2002, IEEE.