Invited Paper to International Conference on Acoustics (ICA) 2004, Kyoto

Multichannel Acoustic Signal Processing for Human/Machine Interfaces - Fundamental Problems and Recent Advances

Walter Kellermann, Herbert Buchner, Wolfgang Herbordt, and Robert Aichner

Chair for Multimedia Communications and Signal Processing
University Erlangen–Nuremberg, Germany
{wk,buchner,herbordt,aichner}@LNT.de

Abstract

Multichannel signal processing techniques for reproduction and acquisition of audio and speech signals at the acoustic human/machine interface offer spatial selectivity and diversity as additional degrees of freedom over single-channel schemes.

In this contribution, we identify the fundamental problems for acquisition and reproduction with distant sources/ listeners as signal separation problems or system identification problems of varying difficulty depending on the available reference information. We analyze the structure of the respective problems and discuss possible solutions. As examples for recent advances in this field, we emphasize speech acquisition systems, and highlight multi-channel acoustic echo cancellation, adaptive beamforming, and blind source separation. The presented algorithms are mainly characterized by their ability to cope well with the nonstationarity of the involved signals and the time-variance as well as the complexity of the acoustic systems, and thereby represent robust solutions to real-world scenarios.

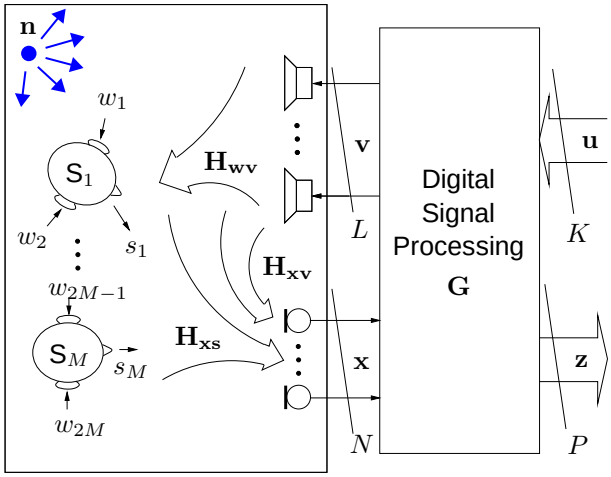

Figure 1: Multichannel acoustic human/machine interface

1. Introduction

We consider an acoustic human/machine interface according to Fig.1 using multiple channels both for reproduction and acquisition of sound, which in general should serve multiple mobile sources and listeners.

For sound reproduction, vector $\mathbf{v}$ contains L loudspeaker signals, which are derived from K source signals captured by vector $\mathbf{u}$. Vector $\mathbf{w}$ of length $2M$ describes the signals at the ears of M listeners, which in the ideal case correspond to a set of desired signals $\mathbf{w}_\mathrm{d}$.

Regarding signal acquisition, vector $\mathbf{s}$ represents M source signals s_i of potential interest. $\mathbf{n}$ captures the noise sources which lead to additive noise vectors $\mathbf{n_w}$, $\mathbf{n_x}$ at the listeners' ears and the microphones, respectively. The objective of signal acquisition is to extract a vector $\mathbf{z}$ from N microphone signals described by vector $\mathbf{x}$ such that, ideally, $\mathbf{z}$ contains $P \leq M$ desired source signals s_i. The matrices $\mathbf{H_{wv}}$,$\mathbf{H_{xv}}$, $\mathbf{H_{xs}}$ describe the transfer characteristics between the respective vector elements. As an essential feature of our scenario, we assume that loudspeakers and microphones need not be close to the human users, and that, ideally, the users should be allowed to move freely. With this general setup, we capture numerous realistic scenarios where natural and synthetic acoustic scenes should be reproduced and/or sources should be recorded for storage, transmission, processing, or interpretation. This includes hands-free speech communication devices in cars, multimedia terminals, and teleconferencing equipment, but also telepresence systems, home theatres, and virtual reality environments. Moreover, such seamless human/machine interfaces are of special importance for user-friendly distant-talking speech recognition and speech dialog systems.

In the following we review the fundamental signal processing problems for creating the desired listeners' signals $\mathbf{w}_\mathrm{d}$ and for extracting the desired source signals $\mathbf{z}$ from $\mathbf{x}$, discuss current solutions, and highlight some examples for recent advances.

2. Fundamental signal processing problems

For the following representation in the discrete time domain we assume that all components of our scenario act as linear, but generally time-variant systems on the defined signals. This allows us to capture the signal processing by a matrix $\mathbf{G}$ representing a linear MIMO ('multiple input/multiple output') system, which realizes linear convolutions on the time domain signals u_i, x_j ($i = 1,\ldots,K; j = 1,\ldots,N$). With submatrices $\mathbf{G_{vu}}, \mathbf{G_{vx}}, \mathbf{G_{zu}}, \mathbf{G_{zx}}$ this reads [1]:

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

The actual ear signals $\mathbf{w}$ and the microphone signals $\mathbf{x}$ are determined by the acoustic environment as follows:

$$\mathbf{w} = \mathbf{H_{wv}} * \mathbf{v} + \mathbf{n_w}, \tag{2}$$

$$\mathbf{x} = \mathbf{H_{xs}} * \mathbf{s} + \mathbf{H_{xv}} * \mathbf{v} + \mathbf{n_x}. \tag{3}$$

We emphasize here that the elements of the matrices $\mathbf{H}_{(\cdot)}$ are commonly impulse responses with a duration of several hundred milliseconds, which are typically modelled by digital FIR filters using around 1000 or more coefficients [2]. With regard to inversion of these systems it must be considered that most of their zeroes lie very close to the unit circle, which causes even much longer impulse responses for the inverse models.

Based on this signal model, we derive now the requirements for the signal processing matrix $\mathbf{G}$. Thereby, we may safely assume that the source signals s_i, the vector of reproduction signals $\mathbf{u}$, and the vector of noise signals $\mathbf{n}$ are mutually independent.

2.1. Sound reproduction

With multichannel sound reproduction we aim at desired ear signals $\mathbf{w}_\mathrm{d}$ which fulfill:

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

where the $2M \times K$ matrix $\mathbf{H}_\mathrm{d}$ describes the usually time-variant impulse responses $h_{ij}(k,l)$ from the j-th input $u_j(k)$ and the i-th ear. Therefore, considering Eqs.1,2, we have to meet:

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

This implies two kinds of signal processing problems:

[1] Convolution of the column vector $\mathbf{x}$ with matrix $\mathbf{A}$ written as $\mathbf{y} = \mathbf{A} * \mathbf{x}$ means that the elements $y_i(k)$ of the output vector $\mathbf{y}$ are computed as $y_i(k) = \sum_{j=1}^{N} \sum_{n=-\infty}^{\infty} a_{ij}(k,n) x_j(n)$ (for the elements of $\mathbf{A}$ we have $a_{ij}(k,n) = a_{ij}(k-n)$ if the system is time-invariant). 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 $\delta(k)$ is the unit impulse. For rank-deficient or non-square matrices $\mathbf{A}$, the matrix $\mathbf{A}^{-1}$ represents the pseudoinverse (see [1]).

A. Deconvolution. Matrix $\mathbf{G_{vu}}$ has to equalize the influence of the room impulse responses $\mathbf{H_{wv}}$ if the signal processing for reproduction should be independent of a given signal vector $\mathbf{u}$:

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

Aside from assuring causality of $\mathbf{G_{vu}}$ by inserting a proper delay into $\mathbf{H}_\mathrm{d}$, the main problem is obviously that, without reference signals at the ears, the matrix $\mathbf{H_{wv}}$ cannot be identified and therefore cannot be inverted. (Note that this problem is even more difficult than the common blind deconvolution problems where channels are identified on the basis of observations at the output of the channels only.) Our problem can only be solved if $\mathbf{H_{wv}}$ can be sufficiently well modelled or measured in advance. Actually, for the latter case, it is known that multichannel systems can efficiently relieve the problem of inverting systems with zeroes close to the unit circle of the individual transfer functions [3], so that $\mathbf{H_{wv}^{-1}}$ can be identified and realized efficiently if observations at the positions of the listeners' ears are available.

B. Noise compensation. From the microphone signals $\mathbf{x}$, we have to extract reference information on the noise and interference signals at the ear, which then can be used for compensation:

$$\mathbf{H_{wv}} * \mathbf{G_{vx}} * \mathbf{x} + \mathbf{n_w} \stackrel{!}{=} \mathbf{0}. \tag{7}$$

This presumes that the noise components at the ears $\mathbf{n_w}$ are completely observable in the microphone signals $\mathbf{x}$, such that a noise vector

$$\mathbf{n_x} = \mathbf{H_{xw}} * \mathbf{n_w} \tag{8}$$

can be observed. From that, one has to form a compensating signal that can be emitted from the loudspeakers and fulfills:

$$\mathbf{H_{wv}} * \mathbf{G_{vx}} * \mathbf{n_x} = \mathbf{H_{wv}} * \mathbf{G_{vx}} * \mathbf{H_{xw}} * \mathbf{n_w} \stackrel{!}{=} -\mathbf{n_w}. \tag{9}$$

Aside from the difficulty of extracting $\mathbf{n_x}$ from $\mathbf{x}$, for signal-independent noise cancellation, we also have to require that the matrix $\mathbf{G_{vx}}$ meets

$$\mathbf{G_{vx}} \stackrel{!}{=} -\mathbf{H_{wv}^{-1}} * \mathbf{H_{xw}^{-1}}. \tag{10}$$

From this we see that $\mathbf{G_{vx}}$ can only be causal if $\mathbf{H_{xw}^{-1}}$ compensates for the acausality of $\mathbf{H_{wv}^{-1}}$. This requires that $\mathbf{H_{xw}}$ is anticausal, i.e., the noise sources are geometrically closer to the reference microphones than to the region of compensation (unless the noise is periodic with some period k_0, $\mathbf{n_w}(k) = \mathbf{n_w}(k - k_0)$). Therefore, in practice, noise compensation calls for distributing many

microphones in the acoustic environment as potential reference sensors.

While the inversion of $\mathbf{H_{wv}}$ in A. is identified as a blind deconvolution problem with unknown output $\mathbf{w}$, the inversion of $\mathbf{H_{xw}}$ in B. is a blind deconvolution problem where the input of the unknown system, $\mathbf{n_w}$, cannot be observed. Note that Eq.7 represents a multichannel system for active noise cancellation ('active noise control') [4]. However, unlike in common active noise cancellation setups, in our scenario we explicitly allow for relatively large distances between actuators and sensors on the one hand, and the spatial region where noise must be compensated on the other hand. So far, no noise compensation schemes for this scenario of hands-free human/machine interfaces are known to the authors.

Standard techniques for sound reproduction do not solve the above problems of deconvolution and noise compensation: With stereo or other multichannel reproduction schemes, the local acoustic environment (represented by $\mathbf{H_{wv}}, \mathbf{n_w}$) is not taken into account and the matrix $\mathbf{G_{vu}}$ is usually a diagonal matrix with (possibly delayed) scalar gain factors, so that the desired listening experience can be provided only in a prescribed sweet spot in an anechoic room without noise. With wavefield synthesis [5] this sweet spot can be extended to an entire plane if a closed contour is sufficiently densely sampled by loudspeakers (e.g., $L = 24, \ldots, 128$). Here, impulse responses in $\mathbf{G_{vu}}$ for auralization of virtual acoustic environments (still without accounting for the local acoustic environment) are common. Current research in wavefield synthesis aims at compensation of the room environment, i.e., at identifying $\mathbf{H_{wv}}$ by solving Eq.6 for an entire region (including the potential positions of the listeners' ears) using off-line measurements [6]. This can be expected to work in a limited frequency range and in idealized environments. However, the impact of the presence of potentially moving persons and their head-related transfer functions is not yet accounted for by this method.

2.2. Signal acquisition

The objective of signal acquisition is a vector $\mathbf{z}$ containing P out of the M original source signals $z_i(k) = s_j(k) * \delta(k - k_0) = s_j(k - k_0), (i = 1, \ldots, P; j \in \{1, \ldots, M\})$, where the delay $k_0 \geq 0$ is required for causal signal processing. For extracting any of the source signals from $\mathbf{x}$, Eq.3 requires that other desired sources and undesired local noise components have to be suppressed, echoes of the loudspeaker signals have to be compensated, and echoes and reverberation of the desired source signal $s_j(k)$ have to be removed from the microphone signals.

For notational simplicity, we assume in the following $P = M$ and an unpermuted mapping of the desired sources $s_j(k)$ to the desired output, $z_j(k)$, so that we obtain as requirement for ideal signal acquisition from Eq.1 with Eq.3:

$$\begin{aligned}
\mathbf{z} &= \mathbf{G_{zu}} * \mathbf{u} + \mathbf{G_{zx}} * \mathbf{x} \\
&= \mathbf{G_{zu}} * \mathbf{u} + \mathbf{G_{zx}} * (\mathbf{H_{xs}} * \mathbf{s} + \mathbf{H_{xv}} * \mathbf{v} + \mathbf{n_x}) \\
&= (\mathbf{G_{zu}} + \mathbf{G_{zx}} * \mathbf{H_{xv}} * \mathbf{G_{vu}}) * \mathbf{u} + \mathbf{G_{zx}} * \mathbf{n_x} \\
&\quad + \mathbf{G_{zx}} * \mathbf{H_{xs}} * \mathbf{s} \\
&\stackrel{!}{=} \mathbf{s} * \delta(k - k_0). \qquad (11)
\end{aligned}$$

This implies three tasks for digital signal processing:

A. Echo cancellation. For compensating the feedback of $\mathbf{u}$ into the output signals $\mathbf{z}$, we obviously have to ensure

$$(\mathbf{G_{zu}} + \mathbf{G_{zx}} * \mathbf{H_{xv}} * \mathbf{G_{vu}}) * \mathbf{u} = \mathbf{0}. \qquad (12)$$

If perfect echo cancellation should be guaranteed independently of the signals $\mathbf{u}$, then

$$\mathbf{G_{zu}} = -\mathbf{G_{zx}} * \mathbf{H_{xv}} * \mathbf{G_{vu}} \qquad (13)$$

must hold. This corresponds to a multichannel version of the classical system identification problem where input and output of the unknown system can be observed. Note that actually only the matrix $\mathbf{H_{xv}}$ describing the acoustic paths from the loudspeakers to the microphones must be identified.

B. Noise suppression. For perfectly suppressing local noise and interference

$$\mathbf{G_{zx}} * \mathbf{n_x} = \mathbf{0} \qquad (14)$$

must be realized. Signal-independent solutions would require $\mathbf{G_{zx}} = \mathbf{0}$, which would prevent the acquisition of any desired signal. Therefore, noise suppression can only be performed without impairment of the desired signals if the noise components in $\mathbf{x}$ can be perfectly separated from the desired signal components, before they are suppressed.

C. Source separation and dereverberation. Assuming that noise and echoes are removed from $\mathbf{x}$, we still have to separate the desired sources and free them from reverberation to obtain

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

This means, for signal-independent solutions, we have to ask for

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

For the elements of the main diagonal of $\mathbf{G_{zx}} * \mathbf{H_{xs}}$ this constitutes a multichannel blind deconvolution problem and for the off-diagonal elements a blind signal separation problem which can also be viewed as an interference cancellation problem similar to Eq.14.

Similarly to the reproduction part, the signal processing subtasks for signal acquisition can be categorized as

problems of either signal separation or system identification. Here, the separation of the components of $\mathbf{x}$ (see Eq.3) is a most crucial part for further identification of $\mathbf{G_{zx}}$, $\mathbf{G_{zu}}$, and $\mathbf{G_{xv}}$: Components correlated with the loudspeaker signals $\mathbf{v}$ must be isolated for identification of the echo cancellers $\mathbf{H_{xv}}$, noise components $\mathbf{n_x}$ should be identified for subsequent suppression, and individual desired source signals must be extracted for immediate use or further processing.

Generally, for separating signal components by multichannel linear signal processing three domains can be exploited: time, frequency, and space. Separation of signal components is relatively simple if the signals are orthogonal in any one of these domains for the given observation interval. For time and frequency, this condition is rarely fulfilled in our scenario: In most cases, noise, interfering signals, and desired signals will overlap at least partially in both time and frequency in the microphone signals $\mathbf{x}$. Fortunately, multichannel signal acquisition also allows spatial filtering to separate signal components originating from different points in space. In reverberant environments, however, the separation of sources according to angles of incidence is also limited, as due to reflections, filtered versions of the source signals may arrive from all angles. As another limitation for signal separation, the sampling theorems have to be observed not only for time and frequency domain but also for spatial apertures [7] to avoid ambiguities, and finite observation intervals will always limit resolution in all three domains. The most critical limitation comes usually from the finite spatial aperture and its sampling by microphones: Audio bandwidths span up to ten octaves which call for many microphones, and for geometrically large apertures at low frequencies.

Among the various system identification tasks in our scenario, echo cancellation is structurally the simplest one, as input and output of the unknown systems can be observed, although the output vector $\mathbf{H_{xv}} * \mathbf{v}$ may be submerged in $\mathbf{x}$. On the other hand, solving the blind deconvolution problem in Eq.16 for realistic scenarios presents a major challenge for current research.

3. Some recent advances in signal acquisition

Rather than attempting a comprehensive overview of this very active research area we present here a synopsis of some recent results with examples from our own work.

3.1. Echo cancellation

For a convenient treatment of the mechanism we assume that $\mathbf{G_{vu}} = \mathbf{I}_{K,K} \cdot \delta(k)$ 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 has been discussed in [8].) Then, Eq.13 reduces to $\mathbf{G_{zu}} = -\mathbf{H_{xv}}$, where the matrices are row vectors with K generally time-variant impulse responses as elements:

$$\mathbf{G_{zu}} = (g_1(k), \ldots, g_K(k)), \quad (17)$$

$$\mathbf{H_{xv}} = (h_1(k), \ldots, h_K(k)). \quad (18)$$

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

$$\widehat{y}(k) = \mathbf{g}^T(k)\mathbf{u}(k), \quad (19)$$

where

$$\mathbf{g}(k) = \left(\mathbf{g}_1^T(k), \ldots, \mathbf{g}_K^T(k)\right)^T, \quad (20)$$

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

with the individual impulse responses and data vectors

$$\mathbf{g}_i(k) = \left(g_{i,0}(k), \ldots, g_{i,L_g-1}(k)\right)^T, \quad (22)$$

$$\mathbf{u}_i(k) = \left(u_i(k), \ldots, u_i(k - L_g + 1)\right)^T, \quad (23)$$

respectively. The estimation error reads:

$$e(k) = y(k) - \widehat{y}(k), \quad (24)$$

where

$$e(k) =: z(k)|_{\mathbf{s}=\mathbf{0}, n_x=0}, \quad y(k) =: x(k)|_{\mathbf{s}=\mathbf{0}, n_x=0}. \quad (25)$$

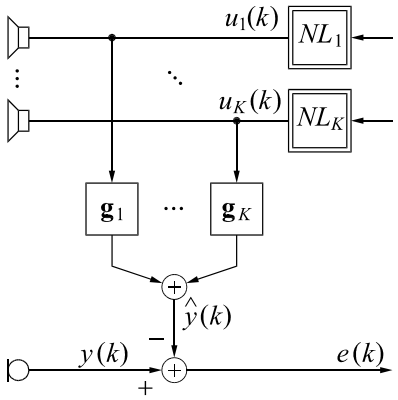

Figure 2: Echo cancellation for K-channel reproduction

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 $\mathbf{g}(k)$:

$$\mathbf{g}(k) = \mathbf{g}(k-1) + \mathbf{k}(k)e(k), \quad (26)$$

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

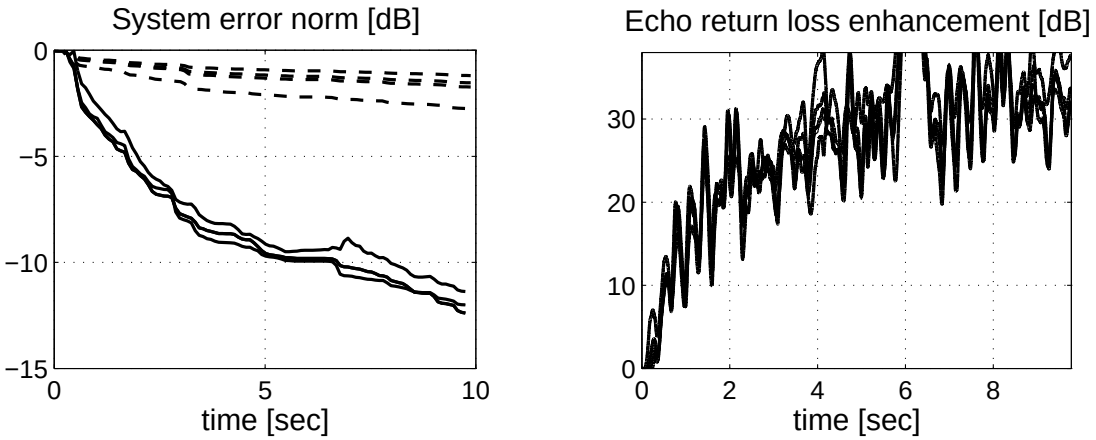

Figure 3: Convergence of DFT-domain adaptation after [9] for music signal reproduction with $K = 2, 3, 4, 5$ channels. System error norm (left) relative to NLMS(dashed lines), echo return loss enhancement (ERLE, right)

[2]) are very popular, for multichannel echo cancellation ($K \geq 2$), algorithms with improved convergence properties are necessary. This is due to the 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 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) [10].

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 $K \cdot L_G = 1000 \ldots 20000$, approximative solutions in the DFT domain are very attractive. In [9] an algorithm is 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 ($\propto \log_{10}(\|\mathbf{G}_{\mathbf{zu}} + \mathbf{H}_{\mathbf{xv}}\|_2^2 / \|\mathbf{H}_{\mathbf{xv}}\|_2^2)$) 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.

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 [11], 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.

3.2. Adaptive beamforming microphone arrays

Beamforming microphone arrays aim at both the signal separation and the suppression of noise and interference, and ideally extract undistorted desired signals. By way of an exemplary design [12], we discuss how these problems can be addressed. A more general treatment of theoretical concepts, alternative approaches, and other aspects of design and applications can be found in [13, 14].

The structure considered here (see Fig.4) is based on a robust version of the Generalized Sidelobe Canceller (GSC) [15] and aims at extracting a single desired signal $z_i \approx s_i$ from $\mathbf{x}$.

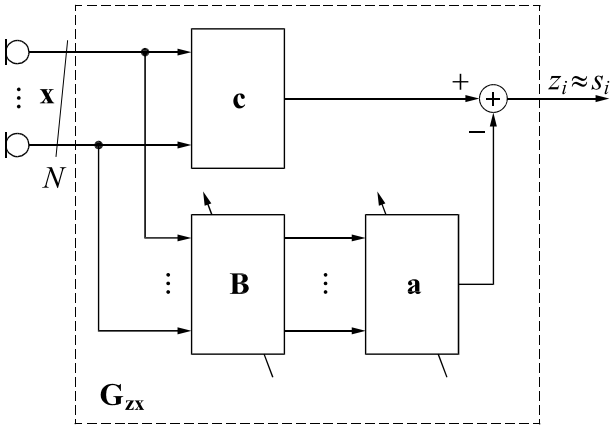

Figure 4: Structure of a robust Generalized Sidelobe Canceller

The GSC principle [16] foresees that a signal-independent beamformer $\mathbf{c}$ filters the sensor signals so that the direct path from the desired source remains undistorted whereas, ideally, other directions should be suppressed. (If necessary, the position of the desired source must be determined by additional localization methods [13].) 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 $\mathbf{B}$. From these, the adaptive interference canceller $\mathbf{a}$ derives an estimate for the remaining noise component in the output of $\mathbf{c}$, by minimizing an estimate of the total output power $E\{z_i^2\}$. Obviously, the fixed beamformer $\mathbf{c}$ and the interference canceller $\mathbf{a}$ jointly perform interference suppression in the sense of Eq.14. 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 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 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 is very beneficial as it allows to completely suppress the signal arriving from the assumed source direction, but it usually cannot 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 [15] 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 [12, 17]. 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).

3.3. Blind source separation

Blind source separation (BSS) aims at separating mixtures of several desired sources, so that $\mathbf{G_{zx}} * \mathbf{H_{xs}} * \mathbf{s} = \mathbf{z} \approx \mathbf{s}$. 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.5 for $M = N = 2$. Blindness also implies that - as opposed to ordinary beamforming - no information on the positions of desired sources is necessary. As such it has been termed 'blind beamforming' [18] and BSS can be understood as realizing a GSC-like structure for each output z_i [19], however, due to the blindness, its components cannot be determined by the same criteria. Lacking reference information, BSS essentially attempts to minimize

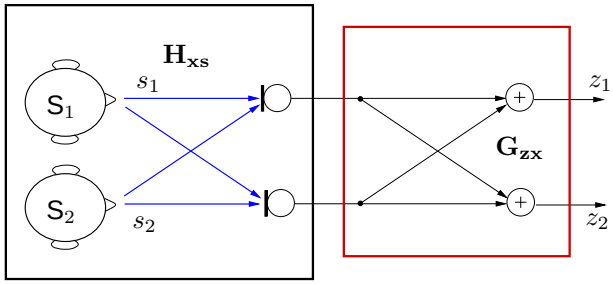

Figure 5: Signal model for BSS

statistical dependency between the output signals, but it should be emphasized that the separation performance of the resulting filters in $\mathbf{G_{zx}}$ is still determined by their spatial selectivity. Note that the optimization criteria do not address 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 $\mathbf{G_{zx}}$:

Nonwhiteness of speech and audio signals can be exploited by simultaneous diagonalization of correlation matrices between $z_i(k), z_j(k-d)$ for several 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., [20]).

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 [21, 22]. Recently, a generic class of algorithms has been presented which simultaneously exploits all three properties and minimizes mutual information [23]. Here, spherical invariant random processes (SIRPs) [24] which represent an efficient model for speech signals if based, e.g., on a Laplacian multivariate probability density function (pdf), can be incorporated into the score function.

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 so that all bins with components of S_i appear at one output of the BSS system, while all bins for S_j appear at the other. Moreover, most frequency domain algorithms are implicitly based on the DFT-inherent circular convolution of the input data in-

stead of the required linear convolution. Heuristic repair mechanisms are common, but 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 [22, 23].

In Fig.6 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. The speech signal mixtures were recorded in a real room with $T_{60} = 0.15$sec at a sampling frequency of 16kHz. Obviously, the HOS-SIRP algorithm [23], which ac-

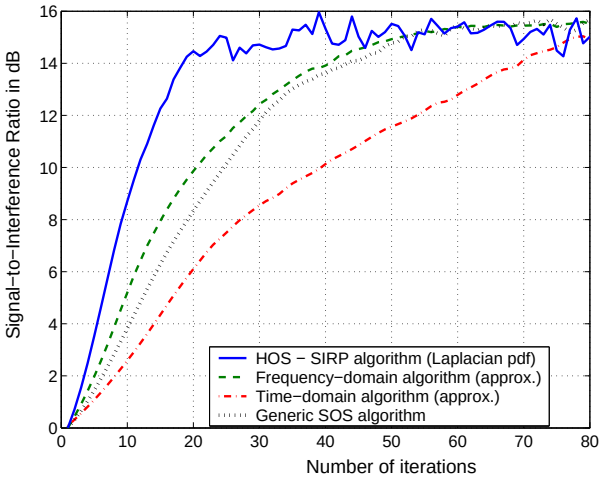

Figure 6: Convergence curves for off-line BSS

counts 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 [21], 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, whereas previously reported real-time implementations of BSS systems all operate in the DFT domain (e.g., [21, 25]).

For future research, robust implementations for $M, N > 2$ and non-square cases ($M \neq N$) present an immediate challenge. Real-world environments call for algorithms that can cope well with diffuse noise, which is known to significantly reduce performance of known algorithms. Finally, a highly attractive avenue of research aims at the extension of the recently found generic BSS algorithms to dereverberation.

4. System integration

For the general scenario of Fig.1 the interaction of the various signal processing components is crucial for the overall performance. Strategies for combining AEC with beamforming and with multichannel reproduction have been discussed in [8] and [26], respectively. As an example for a real system, we consider a DFT-domain implementation of an acoustic front-end for multimedia terminals combining stereo AEC ($K = L = 2$) and robust GSC beamformer ($N = 8$) [12]. Even with simultaneous activity of desired talker and interfering talker, a typical interference suppression of $15dB$ is obtained and loudspeaker echoes can be suppressed by 30dB. The importance of acoustic preprocessing for speech dialogue systems has been verified by measuring word recognition rates for a commercial dictation system ('Dragon System Naturally Speaking Preferred'), see Table 1. It should be mentioned that with greater distance of the talker to the microphone array (and decreasing direct to reverberant signal power ratio), the recognition rates reduce drastically, so that dereverberation will become increasingly important.

Environment	Single mic	FBF	GSC	AEC+ RGSC
Studio (T_{60} : 50ms)	32%	60%	92%	97%
Office (T_{60} : 300ms)	30%	50%	86%	91%

Table 1: Word recognition rates in % for a commercial dictation system (close-talking microphone:= 100%; talker distance to microphone array 0.6m; FBF:= output of fixed beamformer $\mathbf{c}$)

5. Conclusions

Considering the various signal processing problems at the acoustic human/machine interface, it was shown that for signal acquisition, acoustic echo cancellation seems closest to being solved. Noise and interference can also be successfully suppressed in many scenarios by beamforming techniques. Blind source separation works well in low-noise scenarios for two sources. Dereverberation remains a major challenge for the coming years especially with regard to distant-talking speech recognition. On the reproduction side, wavefield synthesis seems to produce satisfactory perceptual audio quality, as long as the influence of the local acoustic environment can be disregarded. Room compensation is under investigation, but wide-band wide-range active noise compensation appears to be out of reach. In summary, it seems safe to conclude that practical relevance and difficulty of the unsolved problems at hand will present many fascinating challenges for digital signal processing on both theoretical and experimental level for the foreseeable future.

6. References

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

[2] 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 Mag.*, vol. 16, no. 4, pp. 42–69, July 1999.

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

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

[5] 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.

[6] S. Spors, A. Kuntz, and R. Rabenstein, "Listening room compensation for wavefield synthesis," in *IEEE Intl. Conf. on Multimedia and Expo (ICME)*, 2003.

[7] D.H. Johnson and D.E. Dudgeon, *Array Signal Processing: Concepts and Techniques*, Prentice Hall, Englewood Cliffs, NJ, 1993.

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

[9] 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., pp. 95-128, Springer, Berlin, 2003.

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

[11] F. Küch, 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. IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing (NSIP'03)*, 2003.

[12] 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.

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

[14] W. Herbordt and W. Kellermann, "Adaptive beamforming for audio signal acquisition," in *Adaptive Signal Processing: Application to Real-World Problems*, J. Benesty, Ed. Springer, Berlin, Jan. 2003.

[15] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive microphone array with improved spatial selectivity and its evaluation in a real environment," in *Proceedings IEEE Intern. Conference on Acoustics, Speech, and Signal Processing*, 1997, IEEE, pp. 367–370.

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

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

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

[19] 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.

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

[21] 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.

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

[23] 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)*, 2003.

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

[25] 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)*, 2003.

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