

Towards Natural Acoustic Interfaces for Automatic Speech Recognition

Walter Kellermann

Multimedia Communications and Signal Processing
University of Erlangen-Nuremberg, Germany

wk@LNT.de

Abstract

Aiming at 'natural' hands-free acoustic human/machine interfaces, the need for according distant-talking automatic speech recognition (ASR) systems increases and presents us with major signal processing challenges at the acoustic front-end. Considering interactive TV as a challenging exemplary application scenario, we investigate the structural problems presented by noisy and reverberant multi-source environments with unpredictable interference and acoustic echoes of loudspeaker signals, and discuss current acoustic signal processing techniques to enhance the input to the actual ASR system. For illustration, the scenario of an EU-funded project on distant-talking interfaces for interactive TV (DICIT) is considered as a challenging example.

1. Introduction

Human/machine interaction became a more and more common part of our everyday life and along with increasingly powerful computational resources the demand for more 'human' interfaces grew continuously. With speech still being the most efficient and natural modality for human-to-human communication, the quest for natural speech as a modality for human/machine interaction persists. While speech dialogue systems with acoustically well-controlled signal acquisition (e.g. via telephones or headsets) are already a commodity, the 'natural' acoustic human/machine interface, which allows the users to be untethered, mobile, and distant from the signal acquisition hardware without the need to wear any extra gear, remains an ambitious goal for research, as the recognition performance of current ASR technology in such scenarios is still very limited.

Generally, ASR systems can try to cope with the challenges due to an uncontrolled acoustic environment on various levels: First, preprocessing of the microphone signals can remove all unwanted signal components and distortion from the desired signal, so that, ideally, clean speech is available for subsequent feature extraction. On a second level, the acoustic model of the ASR system can be adapted to be more tolerant to remaining unwanted signal components and distortions, and, if the word recognition rate is still insufficient for smooth dialogues, language models can correct word recognition errors on a third level. Obviously, remaining errors on the lower levels impair the discriminative performance on higher levels and thereby limit the overall performance. E.g., if the word classifier receives noisy and reverberant signals rather than clean speech, the vocabulary has to be limited and/or the dialogue has to be restricted to well-defined input utterances in order to allow for a satisfactory dialogue.

In this contribution, we concentrate on the acoustic preprocessing in order to deliver a desired signal that can be recognized well by the actual ASR system. We also hint at techniques

which operate on the feature level in order to remove unwanted feature components in the acoustic model of the ASR systems to support classification. After discussing the fundamental problems (partly following [1, 2]), we outline basic techniques and highlight some recent progress in this area. As an illustrative example for many of the unsolved challenges, we refer to "Interactive TV" scenario as it is currently considered in the EU-funded project DICIT [3].

2. The acoustic signal processing scenario

For a generic description of the signal processing scenario, we consider a multiple-input/multiple output (MIMO) system as illustrated in Fig.1, which covers multiple users in an acoustic environment with multichannel sound reproduction and a microphone array for multichannel audio acquisition. On the re-

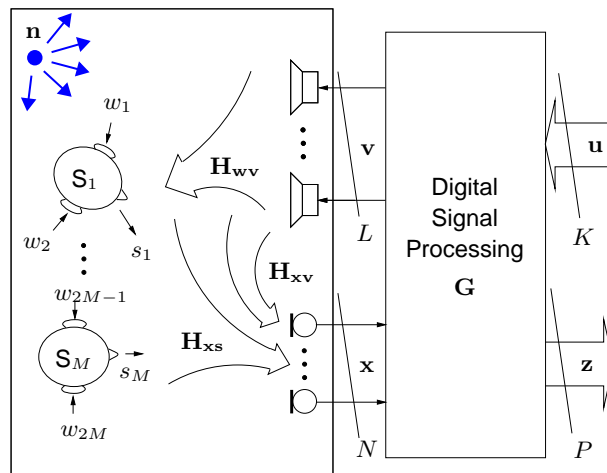


Figure 1: Multichannel acoustic human/machine interface.

production 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}_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 .

The setup in Fig.1 covers a multitude of applications, with varying emphasis and difficulty regarding the individual signal processing problems of the generic scenario. Aside from speech dialogue systems, applications include [1, 2], e.g., hands-

free equipment for telecommunication, teleconferencing, and telecollaboration, seamless interfaces for virtual environments and immersive sound, but also acoustic surveillance, which recently attracted a great deal of attention. Speech dialogue systems by themselves could find their way into an enormously wide range of applications if they allowed for a truly natural voice interaction: For mobile phones, personal digital assistants and mobile computing devices, hands-free operation for communication has been popular for a long time, while voice control is still lagging far behind. Similarly, in cars, hands-free equipment is often an integral part of the user front-end for telephony, but the use of voice dialogues for infrastructure control and navigation systems is still in its infancy. Seamless voice interaction with desktop computers, multimedia terminals, and game stations is another field of applications which still requires major technological advances to exploit its huge market potential. Finally, a class of hands-free applications with potentially even greater user benefit and high demands regarding naturalness of the 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. Here, we concentrate on the typical acoustic scenario for the latter class and refer to Interactive TV as a prototypical scenario, which is also investigated in the EU-funded project DICIT. In this project, all the control and programming information related to a TV system should be accessible via voice in a living room-like environment with multiple users typically several meters away from the microphones attached to the TV set.

Disregarding reproduction techniques and focussing on signal acquisition for natural speech dialogue systems only, the task for the digital signal processing (DSP) unit \mathbf{G} in Fig.1 is simply to extract the desired source signals \mathbf{s} for ASR and to determine the source locations, e.g., to support speaker identification and authentication. Then, obviously, three sources of problems have to be addressed at the acoustic level by the DSP algorithms: Noise, echoes and reverberation of the local source signals \mathbf{s} , and echoes of loudspeaker signals feeding back into the microphones. In the following, we first review the fundamental signal processing problems for signal acquisition, while the main part of the paper is dedicated to recent results with some bias towards work in our own research group.

3. Fundamental problems for signal acquisition

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 ($i = 1, \dots, K; j = 1, \dots, N$). Decomposing \mathbf{G} into submatrices $\mathbf{G}_{\mathbf{v}\mathbf{u}}, \mathbf{G}_{\mathbf{v}\mathbf{x}}, \mathbf{G}_{\mathbf{z}\mathbf{u}}, \mathbf{G}_{\mathbf{z}\mathbf{x}}$, we can write:

$$\begin{pmatrix} \mathbf{v} \\ \mathbf{z} \end{pmatrix} = \mathbf{G} * \begin{pmatrix} \mathbf{u} \\ \mathbf{x} \end{pmatrix} = \begin{pmatrix} \mathbf{G}_{\mathbf{v}\mathbf{u}} & \mathbf{G}_{\mathbf{v}\mathbf{x}} \\ \mathbf{G}_{\mathbf{z}\mathbf{u}} & \mathbf{G}_{\mathbf{z}\mathbf{x}} \end{pmatrix} * \begin{pmatrix} \mathbf{u} \\ \mathbf{x} \end{pmatrix}. \quad (1)$$

The microphone signals \mathbf{x} are given by (see Fig.1)

$$\mathbf{x} = \mathbf{H}_{\mathbf{x}\mathbf{s}} * \mathbf{s} + \mathbf{H}_{\mathbf{x}\mathbf{v}} * \mathbf{v} + \mathbf{x}_n, \quad (2)$$

where we may safely assume that the speech signals s_i and the set of reproduction signals u_i (as contained in \mathbf{v}) are mutually statistically independent and also independent from the

elements of the noise vector \mathbf{x}_n . 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 several hundreds of milliseconds¹. 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, $L_G = f_s \cdot T_{60} / 3$ 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, $f_s = 8\text{kHz}$, $L_G = 1000$ 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 [6]. Moreover, acoustic impulse responses are typically time-variant due to the temperature-dependency of sound velocity and unpredictable changes in the geometric arrangement of scattering objects. Note also that, even if \mathbf{G} represents linear filtering, the elements of \mathbf{G} will usually be determined by complex and usually nonlinear adaptive algorithms.

Based on the system representation given by Eqs.1,2, we analyze now the fundamental problems for signal acquisition to be solved by the signal processing unit \mathbf{G} .

As illustrated by Fig.1, signal acquisition has to extract a vector \mathbf{z} generally containing a subset of P separated and delayed versions of the set of source signals $z_i(k) \approx s_j(k) * \delta(k - k_0)$, ($i = 1, \dots, P; j \in \{1, \dots, M\}$), where the delay $k_0 \geq 0$ is necessary for causality of $\mathbf{G}_{\mathbf{z}\mathbf{x}}$. According to Eq.2, 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.

Generally, it is known that the performance of speech recognizers does not necessarily follow speech quality as measured for human-to-human communication, so that the optimization criteria for removing undesired components for the two cases should ideally be different. For speech recognition an obvious strategy would be to directly optimize the acoustic signal processing with recognition rates as criteria. However, as the susceptibility to different kinds of undesired signal components varies for different speech recognizers and is difficult to predict, the optimization criteria for acoustic signal processing are typically still aiming at minimizing signal power criteria capturing the undesired components and the distortion of the desired signals.

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

$$\mathbf{z} = \mathbf{G}_{\mathbf{z}\mathbf{u}} * \mathbf{u} + \mathbf{G}_{\mathbf{z}\mathbf{x}} * \mathbf{x} \stackrel{!}{=} \mathbf{s} * \delta(k - k_0).$$

To identify the individual signal processing problems, we insert

¹By 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]).

Eq.2 to obtain:

$$\begin{aligned} \mathbf{z} &= \mathbf{G}_{\mathbf{z}\mathbf{u}} * \mathbf{u} + \mathbf{G}_{\mathbf{z}\mathbf{x}} * (\mathbf{H}_{\mathbf{x}\mathbf{s}} * \mathbf{s} + \mathbf{H}_{\mathbf{x}\mathbf{v}} * \mathbf{v} + \mathbf{x}_n) \\ &= (\mathbf{G}_{\mathbf{z}\mathbf{u}} + \mathbf{G}_{\mathbf{z}\mathbf{x}} * \mathbf{H}_{\mathbf{x}\mathbf{v}} * \mathbf{G}_{\mathbf{v}\mathbf{u}}) * \mathbf{u} + \mathbf{G}_{\mathbf{z}\mathbf{x}} * (\mathbf{H}_{\mathbf{x}\mathbf{s}} * \mathbf{s} + \mathbf{x}_n) \\ &\stackrel{!}{=} \mathbf{s} * \delta(k - k_0). \end{aligned} \quad (3)$$

Note that the decomposition of \mathbf{x} presumes that \mathbf{G} is able to separate the components $\mathbf{H}_{\mathbf{x}\mathbf{s}} * \mathbf{s}$, $\mathbf{H}_{\mathbf{x}\mathbf{v}} * \mathbf{v}$, and \mathbf{x}_n .

From Eq.3, 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

$$(\mathbf{G}_{\mathbf{z}\mathbf{u}} + \mathbf{G}_{\mathbf{z}\mathbf{x}} * \mathbf{H}_{\mathbf{x}\mathbf{v}} * \mathbf{G}_{\mathbf{v}\mathbf{u}}) * \mathbf{u} = \mathbf{0}. \quad (4)$$

In a conventional speech dialogue system, \mathbf{u} would typically represent the system outputs, e.g., prompts. Here, we are always assuming that the users are allowed to talk while \mathbf{u} is nonzero ('barge in'-mode). In an interactive TV or home theatre scenario, \mathbf{u} also includes the reproduction signals of the TV system, e.g., a multichannel movie sound. For signal-independent echo cancellation, $\mathbf{G}_{\mathbf{z}\mathbf{u}}$ needs to fulfill

$$\mathbf{G}_{\mathbf{z}\mathbf{u}} = -\mathbf{G}_{\mathbf{z}\mathbf{x}} * \mathbf{H}_{\mathbf{x}\mathbf{v}} * \mathbf{G}_{\mathbf{v}\mathbf{u}}, \quad (5)$$

which is obviously a MIMO system identification problem with both input \mathbf{u} and output \mathbf{x} being observable. Note that actually only the matrix $\mathbf{H}_{\mathbf{x}\mathbf{v}}$ 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

$$\mathbf{G}_{\mathbf{z}\mathbf{x}} * \mathbf{H}_{\mathbf{x}\mathbf{s}} * \mathbf{s} = \mathbf{s} * \delta(k - k_0) \quad (6)$$

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

$$\mathbf{G}_{\mathbf{z}\mathbf{x}} * \mathbf{H}_{\mathbf{x}\mathbf{s}} = \mathbf{I}_{M,M} \cdot \delta(k - k_0), \quad (7)$$

where $\mathbf{I}_{M,M}$ is the $M \times M$ identity matrix. Therefore, we have for the elements of the main diagonal of $\mathbf{G}_{\mathbf{z}\mathbf{x}} * \mathbf{H}_{\mathbf{x}\mathbf{s}}$ a multichannel blind deconvolution problem, and for the off-diagonal elements we have an interference suppression problem similar to that of Eq.8 below.

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

$$\mathbf{G}_{\mathbf{z}\mathbf{x}} * \mathbf{x}_n = \mathbf{0} \quad (8)$$

must be met. Signal-independent solutions would require $\mathbf{G}_{\mathbf{z}\mathbf{x}} = \mathbf{0}$, which, obviously, would also suppress the desired signals. Thus, Eq.8 actually requires $\mathbf{G}_{\mathbf{z}\mathbf{x}}$ to include a linear signal-dependent signal separation unit, such that only the noise is suppressed while Eqs.5,6 can still be met.

In summary, the above subproblems in acquisition involve essentially system identification problems and signal separation/extraction problems. The separation of the various components in \mathbf{x} , i.e., $\mathbf{H}_{\mathbf{x}\mathbf{s}}\mathbf{s}$, $\mathbf{H}_{\mathbf{x}\mathbf{v}}\mathbf{v}$, and \mathbf{x}_n is not only necessary to suppress the noise signals itself, it is also a necessary precondition for obtaining reference information for the various filter optimization algorithms, i.e., for determining $\mathbf{G}_{\mathbf{z}\mathbf{x}}$, and $\mathbf{G}_{\mathbf{z}\mathbf{u}}$.

With the spatial diversity given 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 sufficiently orthogonal in time/frequency to leave the desired sources undistorted after separation. This constitutes a major advantage of the multichannel approaches over single-channel approaches for acoustic human/machine interfaces. One should note, however, that both in time/frequency and 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.

Localisation. The task of localizing and tracking active desired sources S_i is relevant in speech dialogue systems whenever several sources are active and should be distinguished. Localization information can then be used to assign utterances to persons which is especially relevant if authentication is useful or necessary. As a signal processing problem it is different from signal acquisition and reproduction insofar, as the algorithm's output is not a desired signal resulting from modifying input signals, but position information as derived from analyzing the input signals \mathbf{x} by parameter estimation techniques. Clearly, extraction of the desired signals \mathbf{s} should be beneficial, and it seems obvious that knowledge of $\mathbf{H}_{\mathbf{x}\mathbf{s}}$ should facilitate localization. The according techniques for our scenario will be briefly reviewed when discussing recent advances for signal acquisition below.

4. Basic techniques, recent advances, and remaining challenges

In this section we present a brief synopsis of the main techniques and the state of the art regarding the four classes of problems in signal acquisition introduced above, with some bias towards the work of the author's research group. Thereby we mainly concentrate on techniques operating on the signal level while techniques operating on the feature level of speech recognizers are only briefly considered for completeness.

4.1. Acoustic echo cancellation

In the single input/single output case, this supervised system identification problem is considered to be theoretically solved for some time. However, in practice, the fact that the adaptive filters have to continuously identify hundreds to thousands of coefficients and that the reproduction signals \mathbf{u}_i have to suffice as training signals while local sources (\mathbf{s} and \mathbf{n}) act as interferers, requires fast-converging adaptation algorithms with sophisticated control, which are still considered a research subject by many (see e.g. [7]).

For multichannel echo cancellation, i.e., $K \geq 2$ (while still considering only one microphone, $N = 1$), the system identification problem becomes even more challenging, because then, not only the number of filter coefficients multiplies with the number of loudspeakers, but, more importantly, the usually strong and time-varying correlation between the loudspeaker signals (e.g., for stereo or 5.1 sound reproduction) renders the system identification of the K echo paths an ill-conditioned problem [8]. The demand for fast converging and robust adaptation algorithms for very large numbers of coefficients can be met, e.g., by a DFT-domain algorithm, which essentially re-

places the inversion of one matrix of size $(K \cdot L_G) \times (K \cdot L_G)$ as necessary for a time-domain realization by the inversion of L_G matrices of size $K \times K$ instead of [9], 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) [10].

To reduce the correlation between the reproduction channels u_i . For this, three methods have been investigated: (ideally imperceptible) nonlinearities (e.g., [8, 11]), insertion of additive noise (e.g. by an audio codec, such as MP3 or AAC, [12]), and time-varying allpass filters [13, 14]. Obviously, none of these methods will comply with the quest for perfect reproduction, and will be especially objectionable for large numbers of reproduction channels K , where they need to be applied more severely in order to obtain sufficiently fast convergence. More recently, time-varying allpass filtering exploiting the limited spatial resolution of human hearing was proposed and found to be least objectionable ([15]). A powerful means for further increasing convergence speed and ensuring robustness of the echo suppression is to modify the cost function for the filter optimization following the concept of robust statistics [16, 17].

Considering the performance of such echo cancellation systems for a distant-talking speech dialogue system, typically, even the 5-channel AEC system with $K \cdot L_G > 20000$ converges to about 20 dB of echo attenuation within two seconds [10, 1]. This assures sufficient echo suppression for most of the time if the loudspeakers provide persistent excitation \mathbf{u} (e.g., by broadcasting a TV program as in the interactive scenario) so that the adaptation can continuously follow changes in the acoustic environment. If, however, only occasional system prompts are available as input \mathbf{u} to the filter adaptation and changes in the acoustic environment could not be tracked for some time, echo residuals stemming from the initial parts of these prompts must be expected in the signals \mathbf{z} delivered to the ASR system. Therefore, for typical natural dialogue systems allowing barge-in mode, fast convergence of the echo canceller remains of paramount importance.

Considering the transition from a single microphone ($N = 1$) to microphone arrays with respect to echo cancellation, this is straightforward as long as the microphone array processing $\mathbf{G}_{z\mathbf{x}}$ acts as a time-invariant multiple-input/single-output (MISO) system: Then, the echo cancellation simply has to use the output z of $\mathbf{G}_{z\mathbf{x}}$ instead of the microphone signals \mathbf{x} as reference signal to be freed from echoes. If multiple desired signals z_i should be extracted from the acoustic environment, then acoustic echo cancellation has to be implemented for each of them. Once $\mathbf{G}_{z\mathbf{x}}$ becomes time-variant (e.g., when incorporating adaptive algorithms) the echo cancellation algorithms will generally not be able to follow the time-variance of $\mathbf{G}_{z\mathbf{x}}$ and should then cancel the echoes from the individual microphone signals x_i [18].

While the conventional multichannel AEC configuration is suitable, e.g. for most voice-controlled home theatres, other schemes are desirable for more demanding reproduction environments, where a very large number of channels L such as in wave field synthesis is used. For this, a new echo cancellation concept based on *wave domain adaptive filtering (WDAF)* [19] has been proposed to perform echo cancellation in a transform domain with eigenfunctions of the sound field as basis functions. In this area, current research concentrates on combining large loudspeaker arrays with small microphone arrays and studying the effects of moving objects onto the eigenfunctions and the corresponding adaptive filtering [20].

In some common applications, especially with low-cost loudspeakers and overloaded amplifiers, the linear model for the feedback path $\mathbf{H}_{v\mathbf{x}}$ is not valid any more [21, 22]. In [22], the matrix notation as used so far for linear systems was also extended to incorporate Volterra filters, and an efficient DFT-domain algorithm was presented which allows modelling of loudspeaker nonlinearities by second-order Volterra filters [24]. With its large number of parameters and only sparse excitation of system nonlinearities, nonlinear filter structures like Volterra filters are known to converge very slowly in acoustic environments. The quest for fast convergence and computational efficiency has triggered the development and application of many advanced adaptive filtering schemes for this application, as e.g., power filters [25], diagonal coordinate filters [26], iterated partitioned [27], and, most recently, the 'combination of filters' paradigm [28].

4.2. Signal extraction and interference suppression

In the following we consider several multichannel techniques for determining spatiotemporal filters $\mathbf{G}_{z\mathbf{x}}$ to approximate Eqs.6, 8 and/or 7. Seemingly unrelated at first glance, they pursue the same goals and differ essentially regarding the used reference information and optimization criteria. Beamforming aims at extracting desired sources which implies separating them from others and suppressing interference and noise, so that Eqs.6,8 should be approximated. Blind source separation, while separating the desired source signals, concentrates on an approximation of Eq.6, whereas dereverberation aims at fulfilling the stricter requirement Eq.7, which in the multi-speaker case includes perfect source separation.

4.2.1. Beamforming

Beamforming microphone arrays aim at both signal separation and the suppression of noise and interference, and ideally extract undistorted desired source signals. A general treatment of theoretical concepts and aspects of design and applications can be found, e.g., in [29, 30, 31, 32]. 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 a priori, the position of the desired source must be determined by localization methods as discussed below.

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 & Sum beamformer', DSB). This supposes that the location or at least the direction of arrival (DOA) of the desired source is known. Using filters instead of delays, 'Filter & 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} . In principle, a constant aperture width/wavelength ratio for the entire frequency range of interest is desirable to allow for a frequency-independent spatial resolution. Considering a frequency range from 20Hz to 20kHz this would imply a very large aperture at low frequencies and a very small microphone spacing for high frequencies. As a compromise, nested arrays are often used, where microphone spacings increase from the center towards the outer parts [33]. Especially for use with ASR systems, the beamformers will usually also aim at frequency-independent beamwidth, which ensures that sources which are

positioned slightly off the look direction of the beamformer do not appear lowpass-filtered in \mathbf{z} [34, 35].

In many cases, small apertures are desirable which implies that the beamformers operate as differential ('superdirective') beamformers at low frequencies and become highly sensitive to spatially uncorrelated noise (including calibration errors, sensor and amplifier noise)[36]. The latter can be controlled in the design process while simultaneously approximating frequency-independent beamwidth by a newly proposed constrained optimization procedure [37].

As an alternative to computationally low-cost data-independent beamforming, adaptive data-dependent beamforming is attractive for its efficiency in suppressing point-like interfering sources and diffuse background noise. Here, the spatiotemporal filtering $\mathbf{G}_{\mathbf{z}\mathbf{x}}$ is typically 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 [38]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 [39]. 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. To circumvent the constraint optimization problem, the so-called Generalized Sidelobe Canceller (GSC) has been proposed [40], which will cancel the desired signal, however, if the target direction is not precisely known. A robust adaptive version of the GSC has been developed in [41] and further been refined using a DFT-domain implementation [42, 43, 44]. In such a system, for a linear array of $N = 8$ sensors with spacing of 4cm, more than 20dB of interference suppression with negligible distortion of the desired signal can be obtained in environments with moderate reverberation ($T_{60} = 0.3\text{sec}$)[42].

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

For hands-free speech dialogue front-ends, microphone arrays recently gained considerable popularity. Mostly, simple data-independent Delay&Sum beamformers are used, but also RGSC-type adaptive beamformers have been evaluated [42]. As noted above, for optimum ASR performance, beamformers should obviously use recognition scores as optimization criterion. This, however, leads to a complex error surface where useful optima are hard to find [45].

If $P > 1$ desired sources z_i should be extracted, P beamformers can work in parallel using the same N microphone signals \mathbf{x} [46]. As long as these beamformers are time-invariant, they do not interact and subsequent postprocessing may decide which of them should be used as input for an ASR system. The set of beamformers can also be used to track a moving source by switching from one beamformer output to the next. If acoustic echo cancellation is not applied to all P beamformers at all times, but should follow the source of interest, then it is advisable to consider the beamformer as a time-varying system with P states and store the previously obtained AEC filter coeffi-

²Note that the MMSE criterion and the MVDR criterion are defined for stationary processes and based on statistical averaging whereas for nonstationary processes and real data samples, the criteria must be modified to operate with short-term estimates, thereby offering many variations [30]

icients as initial values for a given beam until this beam is used again [47]. This method proved also successful in the Interactive TV scenario of the DICIT project[3]. However, if several adaptive beamformers involving estimation of statistical quantities for individual sources should operate in parallel, the estimation will suffer from the interference of the remaining sources as long as classical optimization criteria, like Mean Square Error (MSE), are used. This problem is better addressed by criteria minimizing mutual information as shown below.

For future applications it might also 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 [48].

4.2.2. Blind source separation

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 in the original blind source separation (BSS) scenarios, where scalar signal mixtures have to be separated [49], in our scenario, BSS algorithms have to separate convolutive mixtures given by $\mathbf{H}_{\mathbf{x}\mathbf{s}} * \mathbf{s}$, so that the output signals are usually still linearly filtered versions of the original signals. Reaching beyond separating the source signals, dereverberation by blind deconvolution aims at extracting the original desired signals by additionally assuming a source model for the desired signals. For distinction, BSS can be understood as blind beamforming [50], and blind dereverberation 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 $\mathbf{G}_{\mathbf{z}\mathbf{x}} * \mathbf{H}_{\mathbf{x}\mathbf{s}} * \mathbf{s} = \mathbf{z} \approx \mathbf{s}$. Here, the \approx sign allows for an additional filtering of each vector element but not for mixing of the vector elements. Actually, BSS can be seen as an interference cancellation system for each output z_i [51]. However, due to the blindness, $\mathbf{G}_{\mathbf{z}\mathbf{x}}$ 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 $\mathbf{G}_{\mathbf{z}\mathbf{x}}$ is nevertheless determined by the spatial selectivity of $\mathbf{G}_{\mathbf{z}\mathbf{x}}$. Note that the optimization criteria of BSS do not address the dereverberation problem Eq.7, although the spatial selectivity of the resulting $\mathbf{G}_{\mathbf{z}\mathbf{x}}$ 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 $\mathbf{G}_{\mathbf{z}\mathbf{x}}$ [52, 53, 54]:

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 be exploited by higher order statistics (HOS) as used for independent component analysis (see, e.g., [55]). 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 prop-

erties are exploited. Successful systems have been presented that are based on second order statistics (SOS) only, and use nonwhiteness and nonstationarity only [54, 56, 57]. TRINICON has been introduced as a generic framework, which allows to simultaneously exploit all three properties and minimizes mutual information between the source signals [52, 53, 58]. Here, spherical invariant random processes (SIRPs) [59] 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 [60]. However, if separation in different 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 [56, 61, 62]. 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 [52].

In mildly reverberant scenarios with $M = N = P = 2, 3$ sources, SOS-based TRINICON algorithms can suppress interfering sources by about 15 . . . 20dB and converge within one or two seconds even in real-time implementations [53, 63]. HOS algorithms converge faster requiring however more computational effort [58].

Research in BSS strives and, a BSS system for up to $M = N = P = 6$ channels has already been demonstrated in real-life situations [64]. Moreover, noise-robust versions have also been published already [65]. Looking at unconstrained scenarios, two major challenges are still awaiting convincing solutions: The distant-talking scenario with a negative Signal-to-Reverberation power ratio (SRR), where reverberation is stronger than the direct sound, and the underdetermined case, where more sources are active than sensors available, $M > N$.

4.2.3. Blind dereverberation

In order to meet Eq.6, dereverberation has to equalize, i.e. invert, the acoustic MIMO system \mathbf{H}_{xs} by \mathbf{G}_{zx} . In [66] it was shown that in principle this is perfectly possible as long as $M < N$ and the original source signals \mathbf{s} are available (and the transfer functions from a source to all the microphones have no common zeros). In the considered scenarios, however, the source signals are not accessible (otherwise dereverberation would not be necessary), which renders the problem of determining the dereverberating system \mathbf{G}_{zx} a blind one. Clearly, speech signals cannot be assumed to be white noise signals, as it is possible for the sources in blind deconvolution problems in data communications. Fig.2 illustrates that speech dereverberation actually only asks for partial blind deconvolution 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. Therefore, blind dereverberation methods for speech signals rely on a speech source model that aims at separating the filtering effect of the vocal tract from that of the acoustic environment [53, 67]. The TRINICON framework foresees to

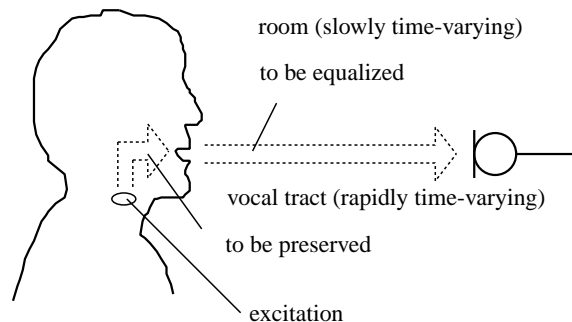


Figure 2: Dereverberation and the PBD principle (from [53]).

incorporate such a speech model into its generic cost function and thereby allows even simultaneous separation and dereverberation for several sources [53]. For $M = N = 2$ an SRR improvement of 7dB without any spurious processing artifacts was reported in [53]. Although further significant progress has been reported (e.g., [68, 69]) over the last few years in dereverberation on the signal level, crucial problems for application in natural environments are still unsolved: So far, most dereverberation algorithms require relatively long observation intervals to determine useful dereverberating filters \mathbf{G}_{zx} , and therefore would not be able to follow potentially fast changes of the acoustic transfer functions in \mathbf{H}_{xs} . Robustness to noise and interference are also not yet addressed by current approaches, so that a robust solution for real-time dereverberation as, e.g., desired for distant-talking speech recognition applications presents still a major challenge for the future.

4.3. Localization

Traditional methods for source localization of sound sources in reverberant rooms follow either one or a combination of the following concepts [70]: a, Steered response beamforming, b, TDOA estimation by crosscorrelation measurement, or c, spectral analysis from array processing techniques. 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 [71]. 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 essentially only be accounted for by tuning window lengths and weighting mechanisms. The main idea of statistical array processing localization techniques 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) [72], 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).

More recently, new concepts have been proposed that explicitly address wideband sources and nonstationary acoustic

environments. Most notably, the adaptive eigenvalue decomposition [73] 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 reverberation in the room. BSS for $M = N = 2$ was shown to achieve the same for two sources simultaneously [74] with even greater robustness to noise. BSS-based localization can be extended to multiple sources in multiple dimensions even for $M = N \geq 3$ as reported in [75, 76].

Aiming at even larger number of wideband sources, the above mentioned array processing methods have recently been applied to signals transformed to the wave domain [77, 78, 79], where the array processing algorithms behave just as for narrowband signals and allow real-time localization of up to $M = 5$ sources.

Beyond estimating instantaneous source locations, tracking of moving sources can be supported by movement models, such as extended Kalman filters [80] or particle filters [81].

4.4. Feature-level processing

While not being in the focus of this contribution, it should not be ignored that for ASR systems the problems of acoustic echoes, noise and interference, and dereverberation can also be tackled at the feature level. The most straightforward approach is to train the acoustic models in the respective environments, so that the recognizers become robust to the undesired signal components (see e.g. [82, 83, 84]). If the acoustic scene at hand, however, does not match the training conditions degradation of recognition performance must be expected. Then two options remain: First, training can try to incorporate a sufficient variety of acoustic scenes. This involves not only a large data collection effort but also implies inferior performance for individual scenes. Second, a parametric model can be extracted from the acoustic scene, which can then be combined with the acoustic model of the ASR system. From the viewpoint of signal processing, the latter approach is more interesting and preferable, especially if a recognizer based on a clean speech model can be used such that it performs well in any acoustic scene without the need for retraining. As a very popular method, cepstral mean subtraction (CMS) is often used to remove the effect of additive noise and linear distortion (e.g., resulting from microphone characteristics) from a feature vector in the MFCC (mel frequency cepstral coefficient) domain. However, this can not model dispersive effects such as reverberation (see [85]). For that, models for the propagation of undesired components over several analysis frames of the feature extraction have to be used [86, 87, 88, 89] with the REMOS framework [85] describing the generic underlying structure for many approaches. Note, that spatial information provided by microphone arrays can not be used on the feature level any more and several simultaneously active sources can also not be separated on the feature level unless very strong assumptions on the source signals are possible (as., e.g., for some musical instruments).

5. Summary and Conclusions

In our discussion of natural seamless acoustic interfaces for human/machine speech dialogues, we considered the various signal processing problems for signal acquisition, which must be solved to obtain high ASR performance in distant-talking real-world acoustic environments. Among those, acoustic echo cancellation as a non-blind MIMO system identification problem,

appears close to being solved, although for multi-channel reproduction system still fundamental problems await elegant solutions. Over the last few years, data-independent as well as adaptive beamforming has reached a certain maturity in achieving the desired signal extraction 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, handling several sources simultaneously. Dereverberation, involving blind deconvolution, will remain a major challenge for distant-talking speech recognition for some more years. Finally, new promising localization techniques for nonstationary wideband sources appear as by-products of blind signal separation and wave-domain signal representation. Feature-level techniques, with their potentially small number of model parameters, promise to be efficient alternatives to signal-level techniques as long as they match the acoustic scenes at hand.

In summary, we may safely conclude that despite of significant progress over the last few years, on the way to a perfect acoustic human/machine interface for natural speech dialogue systems, many fascinating challenges for digital signal processing - on both theoretical and experimental level - remain and can be expected to stimulate intensive for several more years.

6. Acknowledgements

This paper would not have been written without the underlying research contributions of many of the author's PhD students, most notably, Robert Aichner, Herbert Buchner, Wolfgang Herboldt, Fabian KÜch, Anthony Lombard, Armin Sehr, and Heinz Teutsch.

7. References

- [1] Kellermann, W. et al., "Multichannel Acoustic Signal Processing for Human/Machine Interfaces - Fundamental Problems and Recent Advances", Proc. Int. Conf. on Acoustics (ICA), Kyoto, Japan, Apr. 2004
- [2] Kellermann, W., "Acoustic Signal Processing for Next-Generation Human/Machine Interfaces", Proc. of the 8th Int. Conference Digital Audio Effects, Madrid, Spain, Sep. 2005.
- [3] DICIT - Distant talking Interfaces for Control of Interactive TV. EU-Projekt FP6 IST-034624. <http://dicit.itc.it>
- [4] Lawson, C. L. and Hanson, R. J., *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs, NJ, 1974.
- [5] H. Kuttruff, *Room Acoustics*, Spon Press, London, 4th edition, 2000.
- [6] Naylor, P.A., Lin, X., and Khong, A. W. H.. "Near-Common Zeros in Blind Identification of SIMO Acoustic System", Proc. Hands-Free Speech Communication and Microphone Arrays, Seattle, 2008.
- [7] Breining, C. et al., "Acoustic Echo Control". IEEE Signal Processing Magazine 4(1999):42-69.
- [8] Benesty, J., et al., "A Better Understanding and an Improved Solution to the Specific Problems of Stereophonic Acoustic Echo Cancellation." IEEE Trans. on Speech and Audio Processing 6(1998):156-165.

- [9] Buchner, H., Benesty, J., and Kellermann, W., "Multi-channel 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.
- [10] Buchner, H.; Kellermann, W., "Improved Kalman Gain Computation for Multichannel Frequency-Domain Adaptive Filtering and Application to Acoustic Echo Cancellation." Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP'02), 2002, 1909-1912.
- [11] Morgan, D.R.; Hall, J.L.; Benesty, J.: Investigation of Several Types of Nonlinearities for Use in Stereo Acoustic Echo Cancellation. IEEE Trans. on Speech and Audio Processing 9(2001)6, 686- 696
- [12] Gänsler, T., and Eneroth, P., "Influence of audio coding on stereophonic acoustic echo cancellation," Proc. Intern. Conference on Acoustics, Speech, and Signal Processing (ICASSP'98), Seattle, WA, 1998, 3649–3652.
- [13] Ali, M.: Stereophonic Acoustic Echo Cancellation System Using Time-Varying All-Pass Filtering for Signal Decorrelation Proc. of Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'98), 1998, 3689-3692
- [14] Sugiyama, K., Joncours, Y., and Hirano, A., "A stereo echo canceller with correct echo-path identification on an input sliding technique," IEEE Trans. on Signal Processing, vol. 49, 11(2001).
- [15] Herre, J.; Buchner, H.; Kellermann, W.: Acoustic echo cancellation for surround sound using perceptually motivated convergence enhancement. Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'07), I-17 – I-20, 2007
- [16] Gaensler, T.: A double-talk resistant subband echo canceller. Signal Processing 65(1998)1, 89-101
- [17] Buchner, H., et al.: Robust extended multidelay filter and double-talk-detector for acoustic echo cancellation. IEEE Trans. Audio, Speech, Lang. Proc., 14(2006)5, 1633-1644
- [18] Kellermann, W., "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.
- [19] Buchner, H., Spors, S., and Kellermann, W., "Wave-domain adaptive filtering: Acoustic echo cancellation for full-duplex systems based on wave-field synthesis," Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'04), IV-117 – IV-120, 2004
- [20] Schneider, M. and Kellermann, W., "Consequences of modal aliasing in the for acoustic echo cancellation in the Wave Domain", 157th Meeting of the Acoust. Soc. Am., Portland, OR, USA, May 2009.
- [21] Stenger, A., and Kellermann, W., "Adaptation of a memoryless preprocessor for nonlinear acoustic echo cancelling," Signal Processing, vol. 80, pp. 1747–1760, Sep 2000.
- [22] Kch, F., Kellermann, W., "Nonlinear Acoustic Echo Cancellation", in *Topics in Acoustic Echo and Noise Control*, E. Hnsler and G. Schmidt, Springer, Heidelberg, Germany, pp. 205-257, 2006.
- [23] 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.
- [24] Kuech, F., and W. Kellermann, W., "Partitioned block frequency-domain adaptive second-order volterra filter," IEEE Trans. on Signal Processing, vol. 53, no. 2, pp. 564–575, Feb. 2005.
- [25] Kch, F., Kellermann, W., "Orthogonalized power filters for nonlinear acoustic echo cancellation," Signal Processing, Vol. 86, pp. 1168-1181, Jun. 2006
- [26] Kch, F., Kellermann, W., "A Novel Multidelay Adaptive Algorithm for Volterra Filters in Diagonal Coordinate Representation," Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Montreal, Canada, May 2004.
- [27] Zeller, M., Kellermann, W., "Iterated Coefficient Updates of Partitioned Block Frequency-Domain Second-Order Volterra Filters for Nonlinear AEC," Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Vol. 3, Honolulu, USA (HI), pp. 1425-1428, Apr. 2007.
- [28] Azpicueta Ruiz, L., Zeller, M., Arenas-Garcia, J., Kellermann, W., "Novel Schemes for Nonlinear Acoustic Echo Cancellation Based on Filter Combinations," Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), pp. 193-196, Apr. 2009.
- [29] Brandstein, M.S., and Ward, D., (Eds.), *Microphone Arrays: Signal Processing Techniques and Applications*, Springer, Berlin, 2001.
- [30] Herbordt, W., and Kellermann, W., "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.
- [31] Herbordt, W.: Sound capture for human/machine interfaces - Practical aspects of microphone array signal processing. Lecture Notes in Control and Information Sciences, vol. 315, Springer, Heidelberg, 2005.
- [32] Van Trees, H.L., *Optimum Array Processing*, Wiley, New York, NY, 2002.
- [33] Flanagan, J.L., et al: Computer-steered microphone arrays for sound transduction in large rooms. J. Acoust. Soc. Am., 78(1985)5, 1508-1518.
- [34] Goodwin, M.M., Elko, G.W., "Constant beamwidth beamforming," Proc. Intl. Conf. Acoustics, Speech, and Signal Processing, (ICASSP) vol. 1, pp.169-172, 1993.
- [35] Parra, L.C., "Steerable frequency-invariant beamforming," J. Acoust. Soc. Am., 119(2006)6, 3839-3847
- [36] Elko, G., "Differential Microphone Arrays." In Huang, Y.; Benesty, J. (Eds.) *Audio Signal Processing for Next-Generation Multimedia Communication Systems*. Kluwer 2004, 2-65.
- [37] E. Mabande, E., Schad, A., Kellermann, W., "Design of Robust Superdirective Beamformers as a Convex Optimization Problem," Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Taipei, Taiwan, pp. 77-80, Apr. 2009.

- [38] Van Veen, B.D., and Buckley, K.M., "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP magazine*, vol. 5, no. 2, pp. 4–24, Apr. 1988.
- [39] Bitzer, J., and Simmer, K.U., "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.
- [40] Griffiths, L.J., Jim, C.W., "An alternative approach to linear constrained adaptive beamforming," *IEEE Transactions on Antennas and Propagation*, vol. 30, no. 1, pp. 27–34, Jan. 1982.
- [41] Hoshuyama, O., and Sugiyama, A., "A robust adaptive beamformer with a blocking matrix using constrained adaptive filters," in *Proceedings Intern. Conference on Acoustics, Speech, and Signal Processing*. (ICASSP 96), 1996, pp. 925–928.
- [42] Herbordt, W., Buchner, H., and Kellermann, W., "An acoustic human-machine front-end for multimedia applications," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 1, pp. 21–31, Jan. 2003.
- [43] Herbordt, W., Trini, T., and Kellermann, W., "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.
- [44] Herbordt, W., et al.: Multichannel bin-wise robust frequency-domain adaptive filtering and its application to adaptive beamforming. *IEEE Trans. Audio, Speech, Lang. Proc.*, 15(2007)4, 1340-1351
- [45] M. L. Seltzer, B. Raj, R. M. Stern: Likelihood-maximizing beamforming for robust hands-free speech recognition, *IEEE Trans. Speech Audio Process.*, T-SAP-12(5), 489–498, 2004.
- [46] Kellermann, W., "A self-steering digital microphone array," *Proc. Intern. Conference on Acoustics, Speech, and Signal Processing*, Toronto, Ontario, Canada, 1991, pp. 3581–3584.
- [47] Kellermann, W., "Strategies for Combining Acoustic Echo Cancellation and Adaptive Beamforming Microphone Arrays," *Proc. Intern. Conference on Acoustics, Speech, and Signal Processing*, Munich, Germany, 1997, pp.219-222, 1997.
- [48] Herbordt, W., Nakamura, S., Spors, S., Buchner, H., and Kellermann, W., "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.
- [49] Bell, A.J., and Sejnowski, T.J., "An information-maximisation approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 10004–1034, July 1995.
- [50] Cardoso, J.-F., and Souloumiac, A., "Blind beamforming for non-gaussian signals," *IEE Proceedings-F*, vol. 140, no. 6, pp. 362–370, Dec. 1993.
- [51] Araki, S., Makino, S., Hinamoto, Y., Mukai, R., Nishikawa, T., and Saruwatari, H., "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.
- [52] Buchner, H., Aichner, R., and Kellermann, W., "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.
- [53] Buchner, H., Aichner, R., and Kellermann, W., "TRINICON: A versatile framework for multichannel blind signal processing," *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, May 2004.
- [54] Buchner, H., Aichner, R., and Kellermann, W., "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.
- [55] Cardoso, J.-F., "Blind signal separation: Statistical principles," *Proceedings of the IEEE*, vol. 86, no. 10, pp. 2009–2025, Oct. 1998.
- [56] Parra, L., and Fancourt, C., "An adaptive beamforming perspective on convolutive blind source separation," in *Noise Reduction in Speech Applications*, G. Davis, Ed. CRC Press LLC, 2002.
- [57] Buchner, H., Aichner, R., and Kellermann, W., "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.
- [58] Buchner, H., Aichner, R., and Kellermann, W., "Blind source separation for convolutive mixtures exploiting non-gaussianity, nonwhiteness, and nonstationarity," in *Conf. Rec. of the Seventh International Workshop on Acoustic Echo and Noise Control (IWAENC 03)*, Kyoto, Sept. 2003.
- [59] Brehm, H., and Stammer, W., "Description and generation of spherically invariant speech-model signals," *Signal Processing*, vol. 12, pp. 119–141, 1987.
- [60] Sawada, H., Mukai, R., Araki, S., and Makino, S., "Frequency-domain blind source separation," in *Blind Speech Separation*, S. Makino, T.-W. Lee, H. Sawada, Eds., pp. 47–78, Springer, New York, 2007.
- [61] Mukai, R., Sawada, H., Araki, S., and Makino, S., "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.
- [62] Sawada, H., Mukai, R., Araki, S., and Makino, S., "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.
- [63] Buchner, H., Aichner, R., Yan, F., and Kellermann, W., "Real-time convolutive blind source separation based on broadband approach," *Proc. Int. Symposium on Independent Component Analysis and Blind Signal Separation (ICA)*, Sept. 2004.
- [64] Mukai, R., Sawada, H., Araki, S., and Makino, S., "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.

- [65] H., Aichner, Buchner, H., and Kellermann, W., "Convolutional blind source separation for noisy mixtures," Proc. of Joint Meeting of the German and the French Acoustical Societies (CFA/DAGA 2004), Strasbourg, France, March 2004, pp. 583-584.
- [66] Miyoshi, M., Kaneda, Y. "Inverse filtering of room acoustics," *IEEE Trans. Acoustics, Speech, and Signal Processing*, 36(1988)2, 145-152.
- [67] Delcroix, M., Hikichi, T., Miyoshi, M., "Precise dereverberation using multi-channel linear prediction," *IEEE Trans. Acoustic, Speech and Language Processing*, 15(2007)2, 430-440.
- [68] Furuya, K., Kataoka, A., "Hybrid Dereverberation Using Blind Deconvolution and Spectral Subtraction to Compensate for Motion of Source" Conf. Rec. Intl. Workshop on Acoustic Echo and Noise Control (IWAENC), 2006.
- [69] Nakatani, T., Kinoshita, K., Miyoshi, M., "Harmonic-based blind dereverberation for single-channel speech signals," *IEEE Trans. Audio Speech Language Process.*, T-ASLP-15(1) 80-95, Jan. 2007.
- [70] DiBiase, J.H., Silverman, H.F., and Brandstein, M.S., "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.
- [71] Knapp, C.H., and Carter, G.C., "The generalized correlation method for estimation of time-delay," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 24, pp. 320-327, Aug. 1976.
- [72] Krim, H., and Viberg, M., "Two decades of array signal processing research - the parametric approach," *IEEE Signal Processing Magazine*, vol. 5, no. 2, pp. 4-24, Apr. 1988.
- [73] Benesty, J., "Adaptive eigenvalue decomposition for passive source localization," *J. Acoust. Soc. Am.*, vol. 107, no. 1, pp. 384-391, Jan. 2000.
- [74] Buchner, H., Aichner, R., Stenglein, J., Teutsch, H., and Kellermann, W., "Simultaneous localization of multiple sound sources using blind adaptive MIMO filtering," in *Proc. Intern. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Philadelphia, USA, March 2005.
- [75] Lombard, A., Buchner, H., Kellermann, W., "Multidimensional Localization of Multiple Sound Sources Using Blind Adaptive MIMO System Identification." Proc. IEEE Int. Conf. on Multisensor Fusion and Integration for Intelligent Systems (MFI), Sept. 2006.
- [76] Lombard, A., Rosenkranz, T., Buchner, H., Kellermann, W., "Multidimensional localization of multiple sound sources using averaged directivity patterns of blind source separation systems." Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Taipei, Taiwan, pp. 233-236, Apr. 2009.
- [77] Teutsch, H., and Kellermann, W., "EB-ESPRIT: 2D Localization of multiple wideband audio sources using Eigen-Beams," in *Proc. Intern. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Philadelphia, USA, March 2005.
- [78] Teutsch, H., and Kellermann, W., "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.
- [79] Teutsch, H., and Kellermann, W., "Acoustic source detection and localization based on wavefield decomposition using circular microphone arrays," *J. Acoust. Soc. Am.*, Num. 120 (5), Nov. 2006.
- [80] Strobel, N., Spors, S., and Rabenstein, R., "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.
- [81] Ward, D., Lehmann, E., and Williamson, R., "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.
- [82] Huang, X., Acero, A., Hon, H.-W., *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Upper Saddle River, NJ, USA: Prentice Hall, 2001.
- [83] Junqua, J.-C., *Robustness in Automatic Speech Recognition*, Boston, MA: Kluwer Academic Publishers, 1996.
- [84] Stahl, V., Fischer, A., and Bippus, R., "Acoustic synthesis of training data for speech recognition in living-room environments," *Proc. ICASSP '01*, vol. 1, 285-288, Salt Lake City, UT, USA, 2001.
- [85] Sehr, A., and Kellermann, W., "Towards Robust Distant-Talking Automatic Speech Recognition in Reverberant Environments," in *Speech and Audio Processing in Adverse Environments*, E. Hnsler, G. Schmidt (Eds.), Springer, Berlin, 2008.
- [86] Takiguchi, T., Nishimura, M., and Aiki, Y., "Acoustic model adaptation using first-order linear prediction for reverberant speech," *IEICE Trans. Information and Systems*, E89-D(3), 908-914, 2006.
- [87] Hirsch, H.-G., and Finster, H., "A new HMM adaptation approach for the case of a hands-free speech input in reverberant rooms," *Proc. INTERSPEECH '06*, 781-783, Pittsburgh, PA, USA, 2006.
- [88] Raut, C. K., Nishimoto, T., and Sagayama, S., "Model adaptation for long convolutional distortion by maximum likelihood based state filtering approach," *Proc. ICASSP '06*, vol. 1, 1133-1136, Toulouse, France, 2006.
- [89] Sehr, A., and Kellermann, W., "Strategies for modeling reverberant speech in the feature domain," Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), pp. 3725-3728, Apr. 19-24, 2009