[Simon Doclo, Walter Kellermann, Shoji Makino, and Sven Nordholm]

Multichannel Signal Enhancement Algorithms for Assisted Listening Devices

EAR PHOTO—©ISTOCKPHOTO.COM/XRENDER
ASSISTED LISTENING SIGN—© ISTOCKPHOTO.COM/NCANDRE
EARPHONES—IMAGE LICENSED BY INGRAM PUBLISHING

[Exploiting spatial diversity
using multiple microphones]

In everyday environments, we are frequently immersed by unwanted acoustic noise and interference while we want to listen to acoustic signals, most often speech. Technology for assisted listening is then desired to increase the efficiency of speech communication, reduce listener fatigue, or just allow for enjoying undisturbed sounds (e.g., music). For people with normal hearing, assisted listening devices (ALDs) mainly aim to achieve hearing protection or increase listening comfort; however, for hearing-impaired individuals, as the most prominent user group so far, further progress of assisted listening technology is crucial for better inclusion into our world of pervasive acoustic communication.

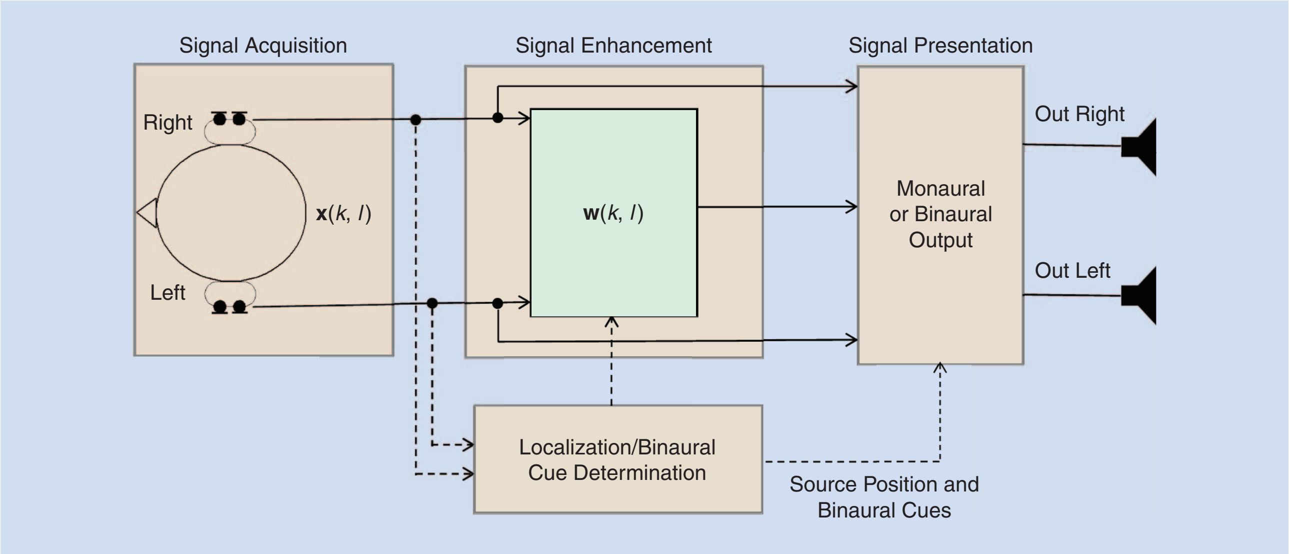
MOTIVATION

The essential functionality of ALDs comprises three steps (see Figure 1): acquiring the signals of interest, enhancing desired and removing undesired components from the acquired signals, and presenting the enhanced signal(s) to the listener.

Given the acquired microphone signals, the efficiency of such devices is largely determined by the performance of the signal processing algorithms for signal enhancement and presentation. Considering that multiple microphones are now common in many

Digital Object Identifier 10.1109/MSP.2014.2366780
Date of publication: 12 February 2015

1053-5888/15©2015 European Union



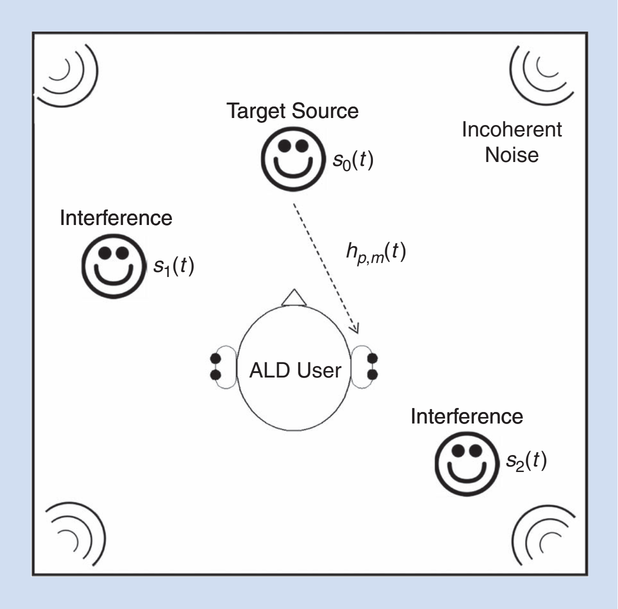
[FIG1] The main processing blocks in an ALD.

listening devices (e.g., hearing aids or mobile phones) and allow to exploit the spatial diversity in addition to the spectrotemporal diversity, multichannel algorithms appear to be decisive for current and future ALDs. Moreover, in contrast to single-microphone signal enhancement algorithms, which have not been shown to improve speech intelligibility but may reduce, e.g., the listening effort, multimicrophone signal enhancement algorithms are capable of increasing speech intelligibility [1], especially when the sound sources have different spatial characteristics.

Although microphone array signal processing, e.g., for teleconferencing systems, is a well-established field dealing with similar problems and signals [2], the problem setting for ALDs exhibits a number of distinctive features. First, the microphone placement is typically constrained by the fact that the devices should be inconspicuously placed at the user's head and should capture the relevant spatial information of the sound sources. Moreover, while all signal enhancement algorithms ideally aim to remove the undesired components and leave the desired components undistorted, the compromises need to be chosen differently depending on the application domain: for ALDs, distortion of the desired signal or annoying noise artifacts will typically be penalized more than a higher level of residual undistorted noise, and the balance between reduced listener fatigue, increased speech intelligibility, and subjective quality plays an even greater role than in other speech communication devices. Finally, for binaural systems that are expected to dominate the future markets, preservation of the critical binaural cues as necessary for a correct spatial perception is crucial [3], not just for the desired signal, but also for the residual noise and interferers.

SCOPE

In this article, we will discuss several algorithms for multimicrophone signal enhancement and presentation that are suitable for ALDs. The considered acoustic scenario is defined by a single source of interest (target source) at any point in time, while multiple interfering point sources (e.g., competing speakers) and additional incoherent noise (e.g., sensor noise, diffuse background noise) may be active simultaneously (see Figure 2). It is assumed that some knowledge is available to distinguish the target source from the interfering sources once they are sufficiently enhanced or separated. Bearing in mind that the wearers of ALDs may move their heads, the relative positions of both the target source as well as the interfering sources must be considered as time-varying, so that source localization and tracking is required.



[FIG2] A scenario with the target source $s_0(t)$, point-like interferers $s_p(t)$, incoherent noise sources, and microphones at the user's head.

The fundamental concept of all considered multimicrophone algorithms relies on spatial and/or spectrotemporal diversity, i.e., the desired components should be separated from the undesired components in the spatial and/or time-frequency domain. The algorithms hence correspond to spatial filtering (often termed *beamforming*) and filtering in the time-frequency domain, respectively. In addition to exploiting the statistics of the available observations, the optimum filter design should also use available prior knowledge, e.g., the estimated or assumed position of the target source. This implies that, in this article, blind source separation (BSS) algorithms [4] are only considered in forms that allow the inclusion of such prior knowledge. Aside from some target-related knowledge, we assume natural unpredictable scenarios that may be arbitrarily complex and time-varying. This implies that the filters must be estimated from currently available observations and cannot be learned in advance, thus algorithms that are based on trained models (e.g., using nonnegative matrix factorization) are not considered in this article. In addition, in time-varying environments, the estimation of the spatial and spectrotemporal information from short observation intervals is of crucial importance, so we will focus on techniques exploiting second-order statistics, keeping the variance of the estimated quantities small.

FOR HEARING-IMPAIRED INDIVIDUALS, AS THE MOST PROMINENT USER GROUP SO FAR, FURTHER PROGRESS OF ASSISTED LISTENING TECHNOLOGY IS CRUCIAL FOR BETTER INCLUSION INTO OUR WORLD OF PERVASIVE ACOUSTIC COMMUNICATION.

SIGNAL MODEL

According to the acoustic scenario in Figure 2, we consider P point sources $s_p(t)$, with t as the discrete time index, in a noise field of unknown coherence, which are recorded by an array of M microphones. The target source is denoted by $s_0(t)$. Assuming the acoustic paths between the sources and the microphones to be linear and time-invariant, the mth microphone signal $x_m(t)$ is given by the convolutive mixing model

$$x_m(t) = \sum_{p=0}^{P-1} h_{p,m}(t) * s_p(t) + n_m(t), \quad m = 1 \ldots M, \qquad (1)$$

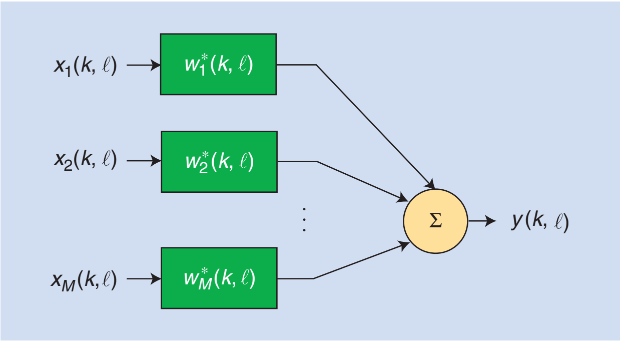
where $n_m(t)$ denotes the noise component in the mth microphone signal, $h_{p,m}(t)$ is the room impulse response (RIR) between the pth source and the mth microphone, and $*$ denotes convolution. Typically, the signals are processed in the short-time Fourier transform (STFT) domain, i.e.,

$$x_m(k,\ell) = \sum_{p=0}^{P-1} h_{p,m}(k) s_p(k,\ell) + n_m(k,\ell), \quad m = 1 \ldots M, \qquad (2)$$

where $x_m(k,\ell)$, $s_p(k,\ell)$ and $n_m(k,\ell)$ denote the STFTs of the respective time-domain signals, with ℓ representing the frame index and k representing the frequency bin index, and where $h_{p,m}(k)$ denotes the acoustic transfer function (ATF) between the pth source and the mth microphone. Note that (2) is strictly speaking only valid for frames that are significantly longer than the RIR length. When this is not the case, a convolutive transfer function model should be used. For conciseness, we omit the dependency on the indices k and ℓ in the remainder of this article. In vector form, the equation set (2) can be written as

$$\mathbf{x} = \mathbf{h}_0 s_0 + \sum_{p=1}^{P-1} \mathbf{h}_p s_p + \mathbf{n} = \mathbf{h}_0 s_0 + \mathbf{v}, \qquad (3)$$

with $\mathbf{x} = [x_1 \ \cdots \ x_M]^T$, and $\mathbf{n}$ and $\mathbf{h}_p$ defined similarly, and $\mathbf{h}_0$ denoting the ATF of the target source. This signal model will form the basis for the subsequent description of the main signal processing tasks with ALDs, i.e., source localization, signal enhancement, and signal presentation.



[FIG3] The filter-and-sum structure.

SIGNAL ACQUISITION

For ALDs in realistic acoustic environments, the ATFs include the microphone characteristics, room acoustics, and filtering effects due to the user's head. The diffraction and reflection properties of the user's head, pinna, and torso are described by the so-called head-related transfer function (HRTF), which is the frequency- and angle-dependent transfer function between a sound source and the user's ear drum in an anechoic environment [5]. The pair of left and right HRTFs contain the so-called binaural cues of a sound source: the interaural time difference (ITD) and the interaural level difference (ILD), which are resulting from the time difference of arrival (TDOA) between both ears and the acoustic head shadow, respectively. In contrast to point sources, the spatial characteristics of incoherent noise can not be properly described by the ITD and ILD, but rather by the interaural coherence (IC) [5]. Binaural cues play a major role in spatial awareness, i.e., for source localization and for determining the spaciousness of auditory objects, and are important for speech intelligibility due to binaural unmasking, e.g., [5].

For capturing the relevant spatial information and binaural cues of the sound sources, in principle, at least two microphones are required, which are preferably mounted on both sides of the head. Ideally, the microphones are placed as close as possible to the corresponding loudspeakers that present the signals to the ear drums to allow the recreation of the authentic spatial impression for the listener. In typical ALDs today, two or three microphones are available on each side of the head, with spacings ranging from 7 mm to 15 mm. Since the positions of the microphones do not coincide with the ear drum, and the acoustic path between the loudspeaker and the ear drum differs from the HRTF, the overall response of the device should be equalized to match the open-ear HRTF [3].

SOURCE LOCALIZATION

The objective of source localization is to estimate the position or the direction of arrival (DOA) of the target source (and possibly the interfering sources), be it for supporting signal extraction or for furnishing signal presentation algorithms with spatial information.

SIGNAL EXTRACTION

The main task is to extract from the given recordings an undistorted version of the target source while all undesired components are suppressed. Two generic approaches can be used to achieve this:

- One can aim at separating all point sources and then pick the target source based on additional knowledge.
- One can directly use the additional knowledge to extract the target source only.

Intuitively, the second approach promises a lower overall algorithmic complexity for a desired performance, as it essentially requires only to separate the target source from all other sources, and obviously avoids the complexity of estimating the potentially large number of irrelevant sources in a given acoustic scene. In addition, the first approach may be limited to setups where the number of microphones is larger than the number of point sources.

Signal extraction is typically achieved using a filter-and-sum structure, depicted in Figure 3, where each microphone signal x_m is passed through a linear filter w_m^* and the outputs are summed. The output signal y is then given in the STFT domain by

$$y = \sum_{m=1}^{M} w_m^* x_m = \mathbf{w}^H \mathbf{x} \tag{4}$$

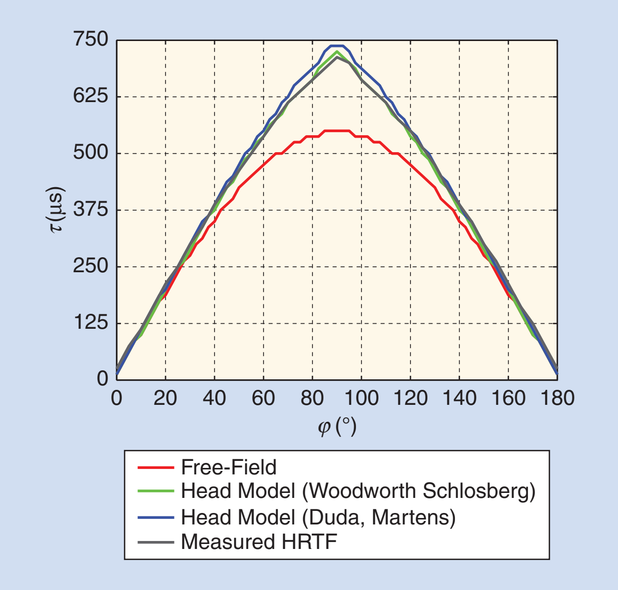
with $\mathbf{w}^H = [w_1^* \; w_2^* \; \ldots \; w_M^*]$. The time-domain output signal may then be computed using the inverse STFT.

While, in principle, additional knowledge may describe source characteristics in both the time-frequency domain or the spatial domain, in this article we will mainly consider additional knowledge in the spatial domain, assuming that the sources are physically located at different positions. Typical prior spatial knowledge is then given by, e.g., the estimated or assumed DOA of the target source relative to the head. With this spatial information, we can support signal extraction algorithms, e.g., a beamformer pointing toward a given DOA or BSS algorithm exploiting the target DOA. These algorithms will be covered in more detail in the sections "Data-Independent Beamforming" and "Statistically Optimum Signal Extraction."

SIGNAL PRESENTATION

After extracting the target source, the enhanced signal is to be presented to the listener, where we need to distinguish between monaural and binaural systems. For a monaural ALD, i.e., a single device on one ear, it seems obvious to just feed the enhanced signal to the loudspeaker of this device. For a binaural ALD, i.e., a system jointly considering and processing the microphone signals of both ears, different signals can be presented to the left and the right ear. This can generate an important binaural advantage since the auditory system can exploit binaural cues and the signal processing algorithms can use information from all microphones on both devices [6, ch. 14]. On the other hand, in a bilateral system where both devices work independently, this potential is not fully exploited since not all microphone signals from both devices are combined. To exploit the full potential of binaural processing, both devices need to cooperate with each other and exchange information or signals, e.g., through a wireless link.

Besides signal extraction, a second major task should be achieved in binaural ALDs: the auditory impression of the acoustic scene, i.e., the spatial perception of the target source, the residual interfering sources and noise, should be preserved. This can be achieved either by so-called binaural rendering of the monaural output signal of the signal extraction algorithm, or by directly incorporating the desired binaural cues into the spatial filter design. These algorithms will be covered in more detail in the section "Presentation of the Enhanced Signals."



[FIG4] TDOAs for different azimuthal directions θ (0° = front, 180° = back) based on free-field assumption, measured HRTFs and two head models, respectively.

SOURCE LOCALIZATION

In principle, any source localization algorithm that can handle multiple nonstationary wideband sources can be used for ALDs [6, ch. 6]. This includes direct methods based on steered-response power (SRP) [2, ch. 8] or subspace methods [Multiple Signal Classification (MUSIC)] [7] and the large and popular class of indirect two-step methods based on TDOA estimation and a subsequent geometric inference of the source position. The latter class comprises cross-correlation-based [8] and cross-relation-based algorithms, e.g., [9] and [10].

The main difference of using these algorithms for ALDs compared to their conventional use results from the fact that the microphones are typically mounted close to the user's head. Therefore, the propagation paths of a point source to the different microphones can not be simply modeled by the free-field TDOA, but the filtering effects of the head should be taken into account. As HRTFs vary between individuals, the results produced by source localization algorithms will always suffer from some uncertainty if the individual HRTFs and the microphone topology are not exactly known. This is especially true for binaural systems, where the relative microphone positions are user dependent and not fixed. However, useful approximations can be employed, which are, e.g., based on spherical head models [11] or measured HRTFs. The TDOAs for different source directions based on the free-field assumption, measured HRTFs, and typical head models is depicted in Figure 4. Alternatively, for binaural systems, computational auditory scene analysis (CASA) algorithms [12] can be used for localizing multiple sources, e.g., incorporating a probabilistic model of the binaural ILD and ITD cues [13].

THE FUNDAMENTAL CONCEPT OF ALL CONSIDERED MULTIMICROPHONE ALGORITHMS RELIES ON SPATIAL AND/OR SPECTROTEMPORAL DIVERSITY.

Given the microphone topology, cross-correlation-based algorithms such as the generalized cross-correlation with phase transform (GCC-PHAT) [8] can be used to localize a single source for ALDs when the head filtering effects are taken into account. However, when multiple sound sources are present, identifying the correct source-specific TDOAs typically becomes very difficult [14]. Generalizations of the GCC, such as SRP-PHAT [2, ch. 8], coherently add up signals originating from a certain point in space to estimate the source likelihood at this position. While conceptually suited for an arbitrary number of microphones and sources, they involve considerable computational complexity for sufficient spatial resolution and are inherently sensitive to reverberation.

More general cross-relation-based algorithms, e.g., [9] and [10], aim at system identification via cross-relation and are naturally suited for identifying relative head-related impulse responses (HRIRs) from the source to the different microphones, delivering TDOA information as long as the direct path can be detected in the identified relative impulse responses. While the adaptive eigenvalue decomposition method in [9] is able to identify relative HRIRs only for a single source while exploiting nonstationarity, the BSS-based method in [10] can robustly localize multiple sources even in noisy and moderately reverberant environments.

Finally, subspace-based source localization algorithms such as MUSIC [7] are in principle also suitable for arbitrary numbers of microphones and sources (assuming the number of sources is known). As they essentially estimate the source positions using the eigenvectors corresponding to the largest eigenvalues of a spatial covariance matrix, the estimates for this covariance matrix must be sufficiently reliable for every frequency bin. Since subspace-based algorithms are separating the signal and noise subspace, where the noise needs to be white or whitened, this is typically difficult to achieve for wideband nonstationary sources in time-varying environments where only short observation intervals can be considered.

DATA-INDEPENDENT BEAMFORMING

A simple but popular way for enhancing the target source in ALDs is data-independent beamforming, where the filters $\mathbf{w}$ in (4) are designed to enhance sources arriving from the (estimated or assumed) target DOA and suppress sources not arriving from this DOA, but do not account for the statistics of the microphone signals. Various data-independent beamformers include delay-and-sum beamformers and superdirective or differential beamformers [2, ch. 2], [15]. For the design of such beamformers, the target DOA and the complete microphone topology need to be known. Data-independent beamformers have mainly been used for monaural devices [16], where robustness against microphone mismatch is crucial due to the closely spaced microphones [17], [18]. For binaural devices, data-independent beamformers have also been proposed, which, however, suffer from spatial aliasing due to the distance between the microphones and require consideration of the head filtering effects, e.g., [19].

STATISTICALLY OPTIMUM SIGNAL EXTRACTION

In contrast to data-independent beamformers, data-dependent signal enhancement methods exploit both the spectrotemporal as well as the spatial information of the microphone signals to extract the target source s_0 (or a filtered version of it) from all interferers and noise [20], possibly equalizing the reverberation effect caused by the ATFs' $\mathbf{h}_0$. Since the filters adapt to the current statistics of the typically nonstationary signals, this will be treated as an optimum multichannel filtering problem in the sequel.

Relying on estimates of either the interference and noise statistics or the target source statistics, two main classes of supervised optimum multichannel filtering will be discussed in the sections "Minimum Variance Distortionless Response Beamformer" and "Multichannel Wiener Filtering." In addition, BSS algorithms, in particular the variants exploiting target-related prior information for constraining the optimization problem to explicitly separate the target source, will be considered in the section "Blind Source Separation." Techniques for estimating the required second-order statistics will be presented in the section "Estimation of Interference and Noise Statistics."

MINIMUM VARIANCE DISTORTIONLESS RESPONSE BEAMFORMER

The minimum variance distortionless response (MVDR) beamformer is a special case of a linearly constrained minimum variance (LCMV) beamformer [20], [21], where the power of the output signal is minimized subject to a single constraint assuring an undistorted response for the target source (or a filtered version of it). Different versions of the MVDR beamformer exist, either using the complete target ATF, the direct path of the ATF, or the relative transfer functions (RTFs). In practice, the MVDR beamformer is often implemented using a so-called generalized sidelobe canceler (GSC) structure [22]–[25].

DERIVATION OF THE MVDR BEAMFORMER

The power spectral density (PSD) of the filter-and-sum beamformer output signal y is given by

$$E\{|y|^2\} = E\{\mathbf{w}^H \mathbf{x}\mathbf{x}^H \mathbf{w}\} = \mathbf{w}^H \boldsymbol{\Phi}_{\mathbf{xx}} \mathbf{w}, \tag{5}$$

where $\boldsymbol{\Phi}_{\mathbf{xx}} \triangleq E\{\mathbf{x}\mathbf{x}^H\}$ denotes the crosspower spectral density matrix of the observed microphone signals. The distortionless response constraint requires that the desired component in the output signal y_{s_0} is equal to the target signal s_0, i.e.,

$$y_{s_0} = \mathbf{w}^H \mathbf{h}_0 s_0 \overset{!}{=} s_0. \tag{6}$$

Hence, by solving the constrained minimization problem

$$\min_{\mathbf{w}} \mathbf{w}^H \boldsymbol{\Phi}_{\mathbf{xx}} \mathbf{w}, \text{ subject to } \mathbf{w}^H \mathbf{h}_0 = 1, \tag{7}$$

we obtain the MVDR filter [20], [21]

$$\mathbf{w}_{\mathrm{MVDR}} = \frac{\boldsymbol{\Phi}_{\mathbf{xx}}^{-1} \mathbf{h}_0}{\mathbf{h}_0^H \boldsymbol{\Phi}_{\mathbf{xx}}^{-1} \mathbf{h}_0}. \tag{8}$$

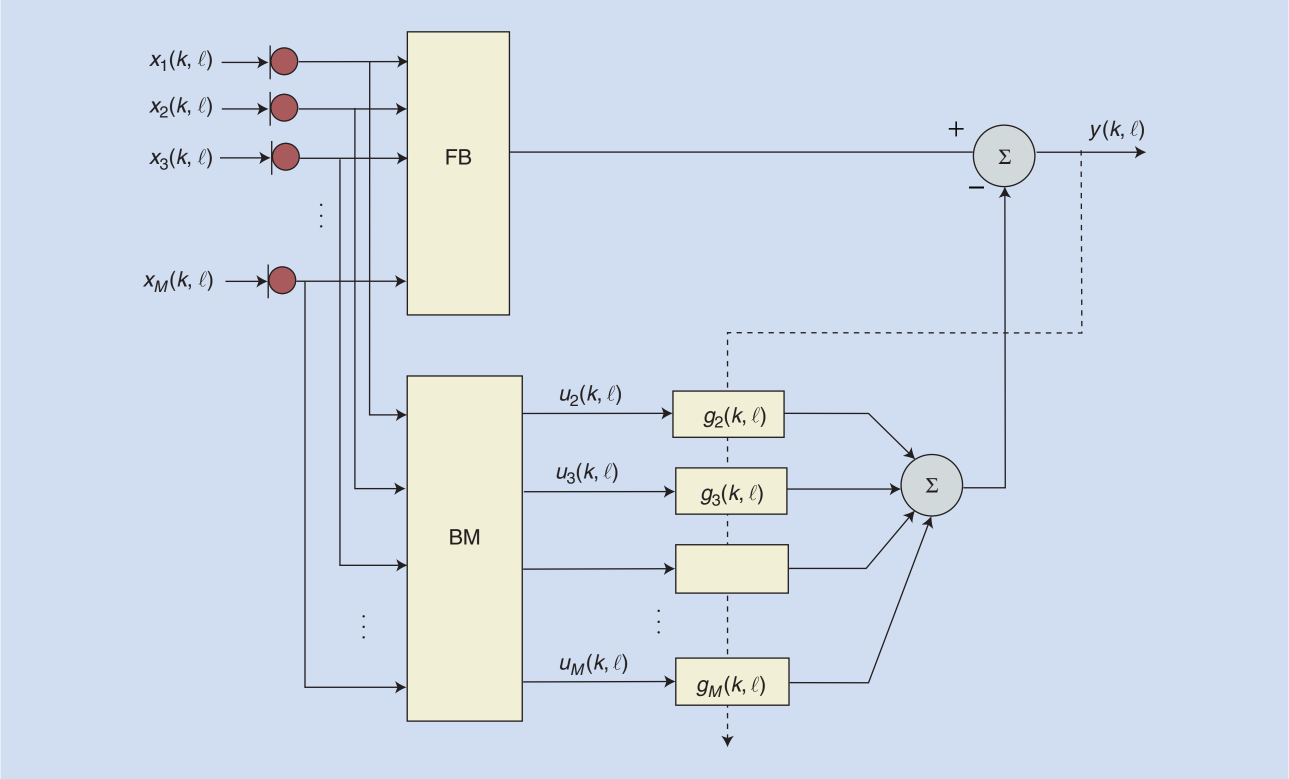
By assuming the target source, the interfering sources and the noise to be mutually uncorrelated and of zero mean, the crosspower spectral density matrix $\boldsymbol{\Phi}_{\mathbf{xx}}$ can be written using (3) as

$$\boldsymbol{\Phi}_{\mathbf{xx}} = \phi_{s_0 s_0} \mathbf{h}_0 \mathbf{h}_0^H + \boldsymbol{\Phi}_{\mathbf{vv}}, \tag{9}$$

where $\boldsymbol{\Phi}_{\mathbf{vv}} \triangleq E\{\mathbf{v}\mathbf{v}^H\}$ denotes the crosspower spectral density matrix of the interference and noise components and $\phi_{s_0 s_0} = E\{|s_0|^2\}$. Using (9), it can be shown that the MVDR filter in (8) can be written as [20]

$$\mathbf{w}_{\mathrm{MVDR}} = \frac{\boldsymbol{\Phi}_{\mathbf{vv}}^{-1} \mathbf{h}_0}{\mathbf{h}_0^H \boldsymbol{\Phi}_{\mathbf{vv}}^{-1} \mathbf{h}_0}. \tag{10}$$

As can be seen, the MVDR filter is solely determined by the crosspower spectral density matrix of the observations and the ATFs $\mathbf{h}_0$. However, due to the high order and the typically time-varying nature of the corresponding RIRs $h_{0,m}(t)$, blindly identifying these impulse responses is generally difficult if at all possible. Hence, instead of using the complete RIRs, one can consider only the direct path of the RIRs (corresponding to the free-field HRIR for the estimated or assumed target DOA), which may, however, lead to target signal distortion, or one can use the so-called RTFs.



[FIG5] The GSC implementation of an MVDR beamformer

MVDR USING RTFs

By constraining the desired component in the output signal to be equal to the speech component at an arbitrarily chosen reference microphone r [24], the constraint in (6) becomes

$$y_{s_0} = \mathbf{w}^H \mathbf{h}_0 s_0 \overset{!}{=} h_{0,r} s_0, \tag{11}$$

which is equivalent to $\mathbf{w}^H \tilde{\mathbf{h}}_0 = 1$, where the RTF $\tilde{\mathbf{h}}_0$ is defined as

$$\tilde{\mathbf{h}}_0 \triangleq \frac{\mathbf{h}_0}{h_{0,r}} = \left[\frac{h_{0,1}}{h_{0,r}} \;\; \frac{h_{0,2}}{h_{0,r}} \;\; \cdots \;\; 1 \;\; \cdots \frac{h_{0,M}}{h_{0,r}}\right]^T. \tag{12}$$

By substituting the ATFs $\mathbf{h}_0$ with the RTFs $\tilde{\mathbf{h}}_0$ in (8) and (10), the modifed MVDR filter is obtained as

$$\tilde{\mathbf{w}}_{\text{MVDR}} = h_{0,r}^* \mathbf{w}_{\text{MVDR}} = \frac{\mathbf{\Phi}_{\text{xx}}^{-1} \tilde{\mathbf{h}}_0}{\tilde{\mathbf{h}}_0^H \mathbf{\Phi}_{\text{xx}}^{-1} \tilde{\mathbf{h}}_0} = \frac{\mathbf{\Phi}_{\text{vv}}^{-1} \tilde{\mathbf{h}}_0}{\tilde{\mathbf{h}}_0^H \mathbf{\Phi}_{\text{vv}}^{-1} \tilde{\mathbf{h}}_0}. \tag{13}$$

Note that blind identification of RTFs is significantly easier than blind identification of ATFs. When noise and interference are absent, this can simply be achieved by dividing the crosspower spectral densities of the microphone signals. When noise and/or interference are present, methods exploiting the nonstationarity of speech or based on the generalized eigenvalue decomposition have been proposed, e.g., [24] and [25].

GSC

The constrained optimization problem of the MVDR beamformer in (7) can be transformed into an unconstrained optimization problem, leading to the highly popular GSC structure [22]–[25], consisting of three main blocks (see Figure 5): 1) a fixed beamformer (FB), ensuring the fulfillment of the constraint in (6) or (11), 2) a blocking matrix (BM), creating so-called noise references u_m, and 3) a multichannel interference canceler g_m, minimizing the residual interference and noise in the output of the FB that is correlated with the noise references. If the target signal leaks into the noise references due to a mismatched BM (e.g., caused by RTF estimation errors or by DOA errors, microphone mismatch, and reverberation when using free-field HRIRs), the target signal will be partially canceled as well. To mitigate this target signal cancellation, the interference canceler is typically adapted only during periods when the target source is inactive; see, e.g., [23]. Moreover, several techniques have been proposed to reduce the speech leakage components in the noise references, e.g., [24], [25], and/or limit the distorting effect of the remaining speech leakage [23], [26], [27], e.g., by imposing a quadratic inequality constraint or by using the so-called speech-distortion-regularized GSC [27].

APPLICATION IN ALDs

The GSC or one of its more robust variants can be considered as the current state-of-the-art solution for monaural hearing devices with an end-fire microphone array configuration, e.g., [28]–[30]. A very popular variant is the adaptive directional microphone (ADM) [15], [28], [29], where the fixed beamformer and the BM are differential beamformers forming a front- and back-oriented cardioid pattern, and an adaptive scalar minimizes the energy arriving from the back hemisphere. A two-microphone implementation was indeed shown to achieve a considerable speech intelligibility improvement for hearing aid users (about 3.4 dB improvement for three babble noise sources) [29].

MULTICHANNEL WIENER FILTER

The second popular class of multichannel signal enhancement techniques is associated with the multichannel Wiener filter (MWF), e.g., [2, ch. 3, 6, 14], [27], [31]. It produces a minimum mean square error (MMSE) estimate of either the target source [2, ch. 3], the speech component at an arbitrarily chosen microphone [2, ch. 6,14], [31], or a reference speech signal [2, ch. 14], [27]. To trade off speech distortion and noise reduction, the so-called speech-distortion-weighted MWF was introduced [27], [31].

Similarly to the MVDR using RTFs, the MWF neither requires a priori information about the microphone configuration nor the position of the target source, making it an appealing approach from a robustness point of view. On the other hand, relying on the second-order statistics of the desired and undesired signal components implies that, for the assumed nonstationary processes, these statistics must be estimated with sufficient accuracy at all times; cf. the section "Estimation of Interference and Noise Statistics."

MMSE ESTIMATION FOR THE MWF

The MWF aims to extract the target source by minimizing the mean square error (MSE) between the (unknown) source signal s_0 and the beamformer output, i.e.,

$$\mathbf{w}_{\text{MWF}} = \underset{\mathbf{w}}{\operatorname{argmin}}\, E\{|s_0 - y|^2\} = \underset{\mathbf{w}}{\operatorname{argmin}}\, E\{|s_0 - \mathbf{w}^H \mathbf{x}|^2\}. \tag{14}$$

Assuming the target source and the interfering sources and noise to be uncorrelated, the solution of (14) is given by

$$\mathbf{w}_{\text{MWF}} = \mathbf{\Phi}_{\text{xx}}^{-1} E\{\mathbf{x} s_0^*\} = \mathbf{\Phi}_{\text{xx}}^{-1} \mathbf{h}_0 \phi_{s_0 s_0}, \tag{15}$$

requiring the ATFs $\mathbf{h}_0$ and the target source PSD $\phi_{s_0 s_0}$ to be estimated, which is a nontrivial task. However, similarly to the MVDR using RTFs, we can also design an MWF aiming at extracting the speech component at an arbitrarily chosen reference microphone r by

$$\tilde{\mathbf{w}}_{\text{MWF}} = \underset{\mathbf{w}}{\operatorname{argmin}}\, E\{|h_{0,r} s_0 - \mathbf{w}^H \mathbf{x}|^2\}, \tag{16}$$

which yields

$$\tilde{\mathbf{w}}_{\text{MWF}} = \mathbf{\Phi}_{\text{xx}}^{-1} E\{\mathbf{x} h_{0,r}^* s_0^*\} = \left(\phi_{s_0 s_0} \mathbf{h}_0 \mathbf{h}_0^H + \mathbf{\Phi}_{\text{vv}}\right)^{-1} \phi_{s_0 s_0} \mathbf{h}_0 h_{0,r}^*. \tag{17}$$

Although it appears that the ATFs and the target source PSD are required to compute (17), the (rank-1) crosspower spectral density matrix $\phi_{s_0 s_0}\mathbf{h}_0\mathbf{h}_0^H$ can be estimated from the second-order statistics of the microphone signals; cf. the section "Estimation of Interference and Noise Statistics."

SPEECH-DISTORTION-WEIGHTED MWF

The MMSE criterion in (16) can be easily generalized to allow for a tradeoff between noise reduction and speech distortion [27], [31] by introducing a weighting factor $\mu \in [0, \infty]$:

$$\tilde{\mathbf{w}}_{\mathrm{SDW}} = \underset{\mathbf{w}}{\operatorname{argmin}}\, E\{|h_{0,r}s_0 - \mathbf{w}^H\mathbf{h}_0 s_0|^2\} + \mu E\{|\mathbf{w}^H\mathbf{v}|^2\}, \quad (18)$$

which is referred to as the speech-distortion-weighted MWF (SDW-MWF). The solution of (18) is given by

$$\tilde{\mathbf{w}}_{\mathrm{SDW}} = (\phi_{s_0 s_0}\mathbf{h}_0\mathbf{h}_0^H + \mu\boldsymbol{\Phi}_{\mathrm{vv}})^{-1}\phi_{s_0 s_0}\mathbf{h}_0 h_{0,r}^{*}. \quad (19)$$

The smaller the factor μ is chosen, the smaller the resulting speech distortion. If $\mu = 1$, the MMSE criterion (16) is obtained. If $\mu > 1$, the residual noise level will be reduced at the expense of increased speech distortion.

RELATIONSHIP BETWEEN MWF AND MVDR

It is interesting to note that the MWF can be decomposed as an MVDR beamformer, exploiting the spatial information of the target and interfering sources, followed by a single-channel Wiener filter (SWF) [2, ch. 3], [32], i.e.,

$$\tilde{\mathbf{w}}_{\mathrm{SDW}} = \underbrace{\frac{\phi_{y_s y_s}}{\phi_{y_s y_s} + \mu\phi_{y_v y_v}}}_{\text{SDW}-\text{SWF postfilter}} \times \underbrace{\frac{\boldsymbol{\Phi}_{\mathrm{vv}}^{-1}\tilde{\mathbf{h}}_0}{\tilde{\mathbf{h}}_0^H\boldsymbol{\Phi}_{\mathrm{vv}}^{-1}\tilde{\mathbf{h}}_0}}_{\text{MVDR beamformer}}, \quad (20)$$

where $\phi_{y_s y_s}$ and $\phi_{y_v y_v}$ denote the PSDs of the desired and undesired components at the output of the MVDR beamformer $\tilde{\mathbf{w}}_{\mathrm{MVDR}}$ using RTFs.

APPLICATION IN ALDs

In [1], a three-microphone MWF implementation for a monaural hearing device was evaluated at different test sites and compared with other single- and multimicrophone noise reduction techniques. In this study it was shown that overall the MWF achieved the largest speech intelligibility improvements (up to 7 dB), even in highly reverberant environments.

BLIND SOURCE SEPARATION

Generalizing the approach of extracting a single desired source, BSS algorithms aim at extracting multiple sources from observed mixtures without requiring prior knowledge on the positions of the sources and the microphones, spatiotemporal signal statistics, or the mixing system. Moreover, they do not need any reference information on the activity of the sources in the spectrotemporal domain. On the other hand, they do require knowledge on the total number of sources and can only separate sources that can be modeled as point sources. Considering time-varying mixing systems, we disregard approaches that perform BSS based on learning from a large amount of data and focus on independent component analysis (ICA)-based methods that are—similar to adaptive filtering approaches—suited to time-varying acoustic scenes [4], [33]–[35].

For the following, we rewrite the STFT signal model in (3) as

$$\mathbf{x} = \sum_{p=0}^{P-1}\mathbf{h}_p s_p + \mathbf{n} = \mathbf{H}\mathbf{s} + \mathbf{n}, \quad (21)$$

describing M noisy observations $\mathbf{x}$ of the convolutive mixture of P point sources s_p. To obtain estimates of the original sources s_p, a linear demixing/separation system $\mathbf{W}$ is applied, consisting of $M \times P$ filters with frequency response w_{mp}, $m = 0, \ldots, M-1$, $p = 0, \ldots, P-1$. The P separated signals y_q, stacked in the vector $\mathbf{y}$, are then obtained as

$$\mathbf{y} = \mathbf{W}^H\mathbf{x} = \mathbf{W}^H\mathbf{H}\mathbf{s} + \mathbf{W}^H\mathbf{n}. \quad (22)$$

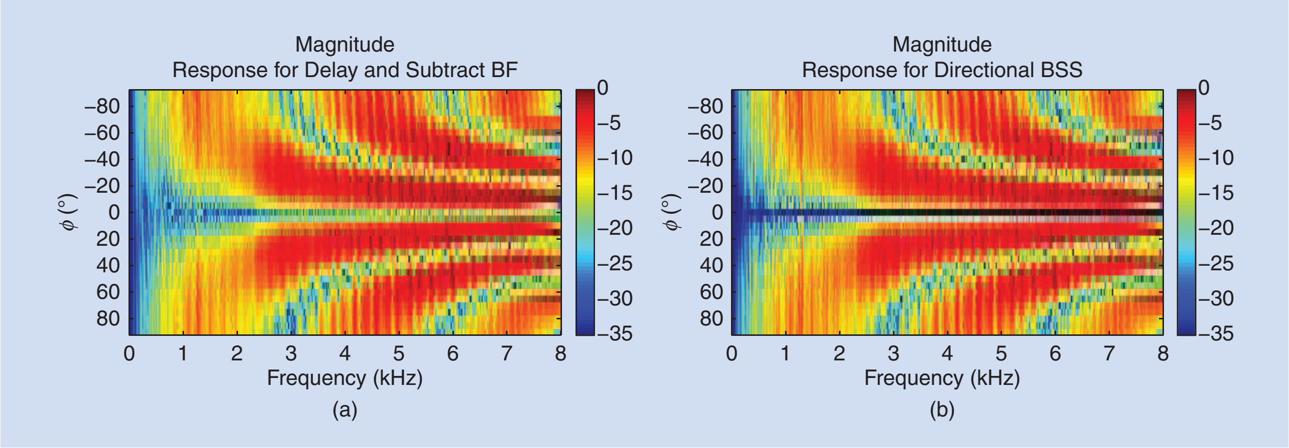
Known methods for identifying optimum demixing filters $\mathbf{W}$ are based on the assumption that the signals to be separated are mutually statistically independent and that enforcing statistically independent outputs y_q of the demixing system yields good estimates of the desired separated source signals s_p. For the mostly assumed case where the number of microphones is larger than or equal to the number of sources ($M \geq P$), an appropriate generic cost function $\mathcal{J}(\ell)$ for frame ℓ, describing an estimate of the Kullback–Leibler divergence between the joint probability density function (pdf) of the output signals y_q and the desired independent outputs, can be formulated as [4, ch. 4]:

$$\mathcal{J}_{\mathrm{ICA}}(\ell) = \sum_{\lambda=0}^{\infty}\beta(\lambda,\ell)\frac{1}{K}\sum_{\kappa=0}^{K-1}\log\frac{\hat{p}_{\mathbf{y},\mathrm{PL}}(\mathbf{y}(\kappa,\lambda))}{\prod_{q=0}^{P-1}\hat{p}_{y_q,L}(y_q(\kappa,\lambda))}, \quad (23)$$

where $\hat{p}_{y_q,L}(y_q(\kappa,\lambda))$ denotes an estimate for the L-variate pdf of a segment of length L of the qth output signal y_q, and $\hat{p}_{\mathbf{y},\mathrm{PL}}(\mathbf{y}(\kappa,\lambda))$ denotes an estimate for the PL-variate joint pdf for all P output signals. Averaging over K frames accounts for the nonstationarity of the data, while the windowing function $\beta(\ell,\lambda)$ describes the weight of a block average at time λ for the cost function at time ℓ, in a similar way as for recursive least squares adaptation. Forming gradients of this cost function, or simplified versions, with respect to the demixing matrix $\mathbf{W}$ allows for maximization of statistical independency with respect to individual data frames (online adaptation, $K = 1, \beta(\lambda,\ell) = 0$ for $\lambda \neq \ell$), as well as for an entire recording (offline adaptation, $K > 1, \beta = \text{constant}$) [35].

It should be noted that using the statistical independence assumption only, the separation system $\mathbf{W}$ can at best be obtained up to a linear filtering uncertainty and a permutation of the outputs, and thus cannot itself identify the inverse mixing system which would solve the deconvolution problem and perfectly dereverberate the source signals [36].

Numerous algorithms have been proposed for ICA of convolutive mixtures, which are often categorized as either time-domain or frequency-domain algorithms. Time-domain algorithms estimate the demixing system $\mathbf{W}$ as finite impulse response filters [35], whereas frequency-domain algorithms



[FIG6] **Interference cancelation in a reverberant environment obtained by (a) null-steering beamformer and (b) ICA ($T_{60} = 300$ ms, interfering point source at 0° at a distance of 1.1 m of a two-microphone array with spacing $d = 15$ cm).**

formulate the demixing problem as a scalar source separation problem for each frequency independently (instantaneous ICA), and implement scalar ICA algorithms for each STFT bin [33], [36], [37].

Similar to adaptive filtering, where time-domain approaches imply a significantly higher computational complexity for obtaining a similar performance as frequency-domain approaches, frequency-domain implementations of ICA (FD-ICA) are computationally more attractive. On the other hand, if these are straightforwardly formulated as independent ICA problems in each STFT bin, the resulting demixing system does not perform a linear but a circular convolution, which is inadequate for demixing a linear mixing system [35]. As an immediate consequence, the so-called internal permutation and scaling problems result: as the outputs of any unconstrained ICA system are only determined up to an unknown scaling factor and a permutation of their order, for FD-ICA the order and the scaling of the outputs may be different for each STFT bin. Therefore the outputs of the scalar ICA units have to be realigned so that for a given output channel y_q all frequency bins belong to the same source [37] and are properly scaled, e.g., by minimizing the average power difference of the outputs y_q relative to the inputs x_m (minimum distortion principle) [38].

In the acoustic signal extraction context, the mechanism of BSS based on ICA has been shown to be equivalent to a set of P adaptive beamformers, each of which aims to extract one source by suppressing all other sources, thereby exploiting the spatial diversity of the microphone signals [39]. Note that for adaptive beamforming, the DOA or the RTFs of the target source should be known, and that it can adapt the required statistics only with given source activity information, while ICA does not need such information.

APPLICATION IN ALDs

To illustrate the spatial filtering capacity of ICA, Figure 6 depicts the overall transfer function $\mathbf{W}^H\mathbf{H}$ from a given source position in a reverberant environment for a null-steering (delay-and-subtract) beamformer and one output channel of an ICA system, thereby demonstrating the actual interference suppression performance in a reverberant environment [40]. The improved spatial null achieved by ICA confirms the hypothesis that, due to capturing all correlated components belonging to the same source in the same output, ICA does not only suppress the direct path but also reflections of an interfering source, e.g., [40]. Nevertheless, one has to bear in mind that the suppression of reflections results from a compromise in the spatial directivity, which a null-steering beamformer cannot offer. Obviously, using the same number of microphones, ICA cannot use more spatial degrees of freedom than a supervised beamformer, and therefore the spatial selectivity of ICA remains limited to what an optimum and ideally controlled beamformer can achieve, as long as it uses the same statistics for determining its parameters [39].

The fact that ICA does not require prior knowledge about source positions, microphone topology, and source activity, and can adapt well during the activity of multiple sources, renders it a highly attractive method for ALDs in complex acoustic environments with unpredictable interference and noise, and usually unknown source and microphone topologies. Unfortunately, however, ICA systems that can robustly and quickly separate more than three sources in real-world environments have not been presented yet, so that scenarios with an unknown number of interferers cannot be handled by such a generic ICA system.

ESTIMATION OF INTERFERENCE AND NOISE STATISTICS

The performance of the signal extraction algorithms discussed in the sections “MVDR Beamformer” and “Multichannel Wiener Filter” critically depends on the estimates of the statistics of the desired and the undesired signal components, respectively. When implementing these algorithms, it is typically assumed that there is a domain where either the desired or the undesired components can be observed alone. While in selected cases, stationarity assumptions may hold reliably to justify a predetermined estimate [41], it must usually be assumed that the statistics of both the desired and the undesired components vary in an unpredictable way and call for instantaneous estimates.

In the spectrotemporal domain, voice activity detection and speech presence probability estimation typically aim at identifying regions in the STFT domain where only undesired components are present, e.g., [6, ch. 5], [42]. Obviously, this is very difficult for the given scenario with interfering speech sources that naturally occupy the same frequency range and whose temporal activity pattern is generally not known, especially if their signal level is comparable to the level of the target source in any of the microphone signals. Therefore, the desired and undesired components can usually only be separated along the time axis. For example, for computing the MWF according to (19), it is typically assumed that the interference and noise can be observed during noise-only periods, so that with the assumed uncorrelatedness of noise and desired speech, the crosspower spectral density matrix $\phi_{s_0s_0}\mathbf{h}_0\mathbf{h}_0^H$ can be estimated as

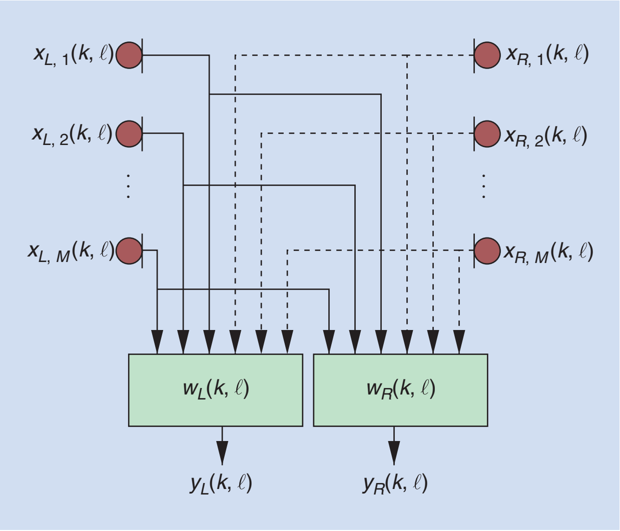
$$\phi_{s_0s_0}\mathbf{h}_0\mathbf{h}_0^H \approx \mathbf{\Phi}_{\mathrm{xx}} - \mathbf{\Phi}_{\mathrm{vv}}, \tag{24}$$

where $\mathbf{\Phi}_{\mathrm{xx}}$ is estimated continually and $\mathbf{\Phi}_{\mathrm{vv}}$ during periods of interference and noise only. As a fundamental problem, however, all these methods still suffer from the fact that the interference and noise estimates cannot be updated while the target source is active, so that they are prone to failure with nonstationary noise and interference, such as human speakers.

On the other hand, in the spatial domain, reference information for all the interference and noise components can be obtained by suppressing the target source. Here, the spatial selectivity allowed by the microphone array topology constitutes the main limitation. Exploiting the spatial domain for obtaining interference and noise reference information is an inherent feature of the GSC (cf. the section "MVDR Beamformer"), where the BM aims to suppress the target source. For moving sources and multipath propagation scenarios, robust adaptation schemes for the BM have already been proposed, e.g., [23]. These concepts still require knowledge about the activity of the target source, as the BM should only be adapted when the target source is dominant. If the DOA of the target source is known, its activity can be monitored by directing both a delay-and-sum beamformer and a delay-and-subtract beamformer in this direction and inferring the activity from the ratio of its output powers, see, e.g., [23]. However, these noise estimates will still be suboptimal if the BM could not be updated while the target source changed its position relative to the microphones on the user's head or the acoustic environment changed.

More recently, a constrained BSS scheme has been proposed to identify the filters of two-channel blocking matrices [40], which does not need source activity information and continuously delivers up-to-date estimates for noise and interference. For this, the cost function in (23) is complemented by a quadratic constraint for one output (here y_p) steering a null toward the target source:

$$\mathcal{J}_{\mathrm{C}}(\mathbf{W}) = \|\mathbf{w}_p^H\mathbf{d}\|_2^2, \tag{25}$$



[FIG7] The general binaural processing scheme.

where $\mathbf{w}_p$ denotes the vector of demixing filters in $\mathbf{W}$ which produce the output y_p, and $\mathbf{d}$ denotes the steering vector corresponding to the DOA of the direct path of the target source. This yields the constrained ICA cost function

$$\mathcal{J}_{\mathrm{C-ICA}}(\mathbf{W}) = \mathcal{J}_{\mathrm{ICA}}(\mathbf{W}) + \eta\mathcal{J}_{\mathrm{C}}(\mathbf{W}), \tag{26}$$

whose minimization suppresses the target in one output channel and thereby provides a reference for all other sources and noise of unknown coherence. The weight is typically chosen as $\eta \approx 0.5\ldots0.8$ with larger values required if interfering sources are close to the target source. It should be noted that, although the constraint captures only the direct path, constrained ICA will intrinsically also aim at suppressing all correlated components, i.e., reflections of the target source, in the same output, thereby providing an advantage over a delay-and-subtract beamformer as shown in Figure 6. As the most attractive advantage, however, the fundamental concept of ICA assures a continuous update of the noise estimate without the need of estimating the activity of the involved sources. Recently, it was also shown that this concept can be generalized to identify all RTFs required for the BM of a GSC with an arbitrary number of constraints [43].

PRESENTATION OF THE ENHANCED SIGNALS

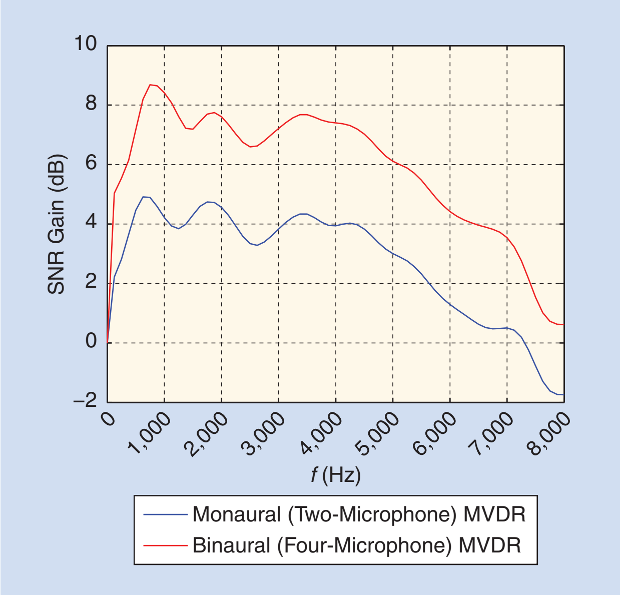
After extracting the target source using data-independent beamforming or statistically optimum filtering (cf. the sections "Data-Independent Beamforming" and "Statistically Optimum Signal Extraction"), the enhanced signal needs to be presented to the listener. While microphone placement is important to maintain a close relationship to the individual HRTFs, we also need to distinguish between a monaural system, i.e., a single device on one ear, and a binaural system, i.e., a system jointly processing signals, at both ears. While for a monaural system it seems obvious to just feed the enhanced signal to the loudspeaker of this device, for a binaural system different output signals y_L and y_R can be generated and presented to the left and the right ear (cf. Figure 7).

In a bilateral system, i.e., a set of two independently operating monaural systems, each device uses its own microphone signals and optimizes its filter coefficients independently, which may lead to a distortion of the binaural cues and hence the localization ability [44]. To achieve true binaural processing, both devices need to cooperate with each other and exchange information or signals, e.g., through a wireless link. Currently, the first commercial systems reach the market which exchange microphone signals in full-duplex mode. These systems pave the way to future implementations of fully fledged binaural multimicrophone signal extraction algorithms, where microphone signals from both devices are processed and combined in each device. The gain in noise reduction performance of a binaural over a monaural system is exemplarily shown for an MVDR beamformer in Figure 8.

TO EXPLOIT THE FULL POTENTIAL OF BINAURAL PROCESSING, BOTH DEVICES NEED TO COOPERATE WITH EACH OTHER AND EXCHANGE INFORMATION OR SIGNALS.

The objective of a binaural speech enhancement algorithm is not only to selectively extract the target source and to suppress interfering sources and background noise, but also to preserve the auditory perception of the complete acoustic scene. This can be achieved by preserving the binaural cues, i.e., ITD, ILD, and IC, of the target source and the residual interfering sources and background noise. In addition to monaural cues, these binaural cues play a major role in spatial awareness and localization and are very important for speech intelligibility due to binaural unmasking, e.g., [5].

All discussed signal enhancement algorithms in the sections "Data-Independent Beamforming" and "Statistically Optimum Signal Extraction" essentially generate a single-channel output signal. Since in a binaural system two output signals (i.e., one for each ear) are required, this single-channel output signal can either be binauralized, e.g., using binaural spectral postfiltering techniques [19], [45], [46] or by mixing the output signal with scaled (noisy) microphone signals [47], [48], or two different complex-valued spatial filters can be optimized, where the desired binaural cues are directly incorporated into the spatial filter design, e.g., [48]–[50]. Although the latter paradigm allows for more degrees of freedom to achieve noise reduction, there is typically a tradeoff between noise reduction performance and binaural cue preservation.



[FIG8] The SNR gain of a monaural and a binaural MVDR beamformer (diffuse noise field).

In binaural spectral postfiltering techniques, the same real-valued gain is applied to one microphone signal of each device, where a gain close to one is applied when the STFT bin should be retained (target source), and a gain close to zero is applied when the STFT bin should be suppressed (interfering source or background noise). This spectral gain can, e.g., be computed by comparing the estimated binaural cues with the expected cues of the target source or based on the temporal fluctuations of the ITD [45]. Other commonly used approaches compute the spectral gain based on the output signal of a data-independent or statistically optimum spatial filter (e.g., MVDR beamformer, BSS) [19], [46]. Although binaural spectral postfiltering techniques preserve the binaural cues of all sound sources, in essence, they can be viewed as single-channel noise reduction techniques, hence typically introducing speech distortion and exhibiting single-channel noise reduction artifacts (e.g., musical noise), especially at low input SNRs.

The MVDR beamformer (using RTFs) and the MWF can be straightforwardly extended into a binaural version producing two output signals, by estimating the speech component in two reference microphone signals, i.e., one on each hearing aid [48]. In [48] and [44], it was shown both analytically and using subjective listening experiments that the binaural MWF preserves the binaural cues of the target source but distorts the binaural cues of interferers and noise, such that all components are perceived as coming from the direction of the target source. Clearly, this is undesired and, in some situations (e.g., traffic), even dangerous. To optimally benefit from binaural unmasking and to optimize the spatial awareness of the hearing aid user, several extensions for the binaural MWF and the MVDR beamformer have been proposed, which aim at also preserving the binaural cues of the residual noise component by including cue preservation terms in the binaural cost function, e.g., [48]–[50]. These include either RTF preservation or interference rejection constraints for directional interfering sources [48], [49], or IC preservation constraints for diffuse noise [50]. Another approach is partial noise estimation, which corresponds to mixing the binaural outputs with scaled versions of the noisy reference microphone signals [48].

APPLICATION IN ALDs

In [30] and [44], the performance of the binaural MWF and some of its extensions has been perceptually evaluated, both in terms of speech intelligibility and localization performance. First, it was shown that the binaural MWF achieved significant speech intelligibility improvements compared to the bilateral MWF and the bilateral ADM. This demonstrates that transmitting and processing microphone signals from both devices can result in a significant gain in noise reduction, especially when multiple interfering sources are present. Second, using a localization experiment in the frontal horizontal hemisphere, it was shown that using the binaural MWF with partial noise estimation it is possible to preserve spatial awareness without significantly affecting speech intelligibility.

SUMMARY AND OUTLOOK

In this article, we have presented an overview of several multimicrophone signal enhancement algorithms for ALDs and have addressed other important issues, such as microphone placement and binaural signal presentation. Using appropriate processing with multiple microphones in a binaural ALD allows both speech intelligibility improvement as well as a preservation of the auditory perception of the acoustic scene.

Future work in this area will focus both on algorithmic aspects and a better integration of psychoacoustics. On the algorithmic side, more accurate and robust estimation and careful exploitation of comprehensive spatiotemporal signal statistics for all relevant sources in highly time-varying scenarios will be necessary to allow for the ultimate desired binaural presentation. The learning of acoustic scenarios and source characteristics can certainly be expected to contribute to reaching this goal. Optimum distribution of the computational load over the available computing hardware via bit rate-constrained "body area networks" will constitute another challenge to algorithm developers. On the psychoacoustic side, ideally, meaningful criteria are desirable that can directly be integrated into the cost functions to allow perceptually optimum signal processing at any given time instant. This may start from incorporating general knowledge about well-known noise masking effects combined with knowledge on the relative importance of certain binaural cues as used already in audio coding and reach to more powerful, yet unknown models for human hearing. For each individual, it should be merged with knowledge about possible hearing impairments or personal listening preferences, i.e., a so-called auditory consumer profile. One may speculate that with suitable user interfaces, the traditional fitting procedures will be replaced by training procedures supervised by the user and even the cost functions for optimizing the multichannel filtering will be as individual as the users themselves. All of these developments will certainly benefit from the integration into handy, but powerful personal computing platforms that are already emerging.

[ALL OF THESE DEVELOPMENTS WILL CERTAINLY BENEFIT FROM THE INTEGRATION INTO HANDY, BUT POWERFUL PERSONAL COMPUTING PLATFORMS THAT ARE ALREADY EMERGING.]

ACKNOWLEDGMENTS

This research was supported in part by the Research Unit FOR 1732 "Individualized Hearing Acoustics" and the Cluster of Excellence 1077 "Hearing4All," funded by the German Research Foundation (DFG), and by the European Union under contract FP6004171 (HEARCOM).

AUTHORS

Simon Doclo (simon.doclo@uni-oldenburg.de) received the Ph.D. degree in applied sciences from Katholieke Universiteit Leuven, Belgium, in 2003. Since 2009, he has been a full professor at the University of Oldenburg, Germany, and a scientific advisor for the Fraunhofer Institute for Digital Media Technology. His research activities center on signal processing for acoustical and biomedical applications. He received the EURASIP Signal Processing Best Paper Award in 2003 and the IEEE Signal Processing Society (SPS) Best Paper Award in 2008. He was member of the IEEE SPS Technical Committee on Audio and Acoustic Signal Processing and is an associate editor of the EURASIP *Journal on Advances in Signal Processing*.

Walter Kellermann (wk@LNT.de) received the Dipl.-Ing. degree in electrical engineering in 1983 and the Dr.-Ing. degree in 1988. From 1989 to 1990, he was a postdoctoral member of technical staff at AT&T Bell Laboratories. He has been a professor of communications at the University of Erlangen-Nuremberg, Germany, since 1999. He (co)authored 16 book chapters and more than 200 refereed papers and was corecipient of several best paper awards for IEEE publications. He served as an associate editor and as a guest editor to various journals and was a general chair for several international conferences. His service to the IEEE Signal Processing Society (SPS) includes Distinguished Lecturer, chair of the IEEE SPS Technical Committee on Audio and Acoustic Signal Processing, and member-at-large for the SPS Board of Governors. He is an IEEE Fellow.

Shoji Makino (maki@tara.tsukuba.ac.jp) received the Ph.D. degree from Tohoku University, Japan, in 1993. Since 2009, he has been a professor at the University of Tsukuba. His research interests include acoustic signal processing for speech and audio applications. He received the ICA Unsupervised Learning Pioneer Award and the IEEE Machine Learning for Signal Processing Competition Award. He has served on the IEEE Signal Processing Society (SPS) Technical Directions Board, Awards Board, and Conference Board. He was associate editor of *IEEE Transactions on Speech and Audio Processing*. He is the chair of the IEEE SPS Technical Committee on Audio and Acoustic Signal Processing. He is an IEEE Distinguished Lecturer and an IEEE Fellow.

Sven Nordholm (s.nordholm@curtin.edu.au) received the Ph.D. degree in signal processing from Lund University, Sweden, in 1992. Since 1999, he has been a full professor at Curtin University, Perth, Australia, and an advisor for Hearmore and Sensear. His research activities focus on signal processing and communication in

acoustic media. He founded Sensear in 2006 and is also its technology inventor. He is a member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing and is an associate editor of *Journal of the Franklin Institute*.

REFERENCES

[1] H. Luts, K. Eneman, J. Wouters, M. Schulte, M. Vormann, M. Buechler, N. Dillier, R. Houben, W. A. Dreschler, M. Froehlich, H. Puder, G. Grimm, V. Hohmann, A. Leijon, A. Lombard, D. Mauler, and A. Spriet, "Multicenter evaluation of signal enhancement algorithms for hearing aids," *J. Acoust. Soc. Amer.*, vol. 127, no. 3, pp. 2054–2063, Mar. 2010.

[2] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*. New York: Springer, 2001.

[3] J. Blauert, *The Technology of Binaural Listening*. New York: Springer, 2013.

[4] S. Makino, T.-W. Lee, and H. Sawada, *Blind Speech Separation*. New York: Springer, 2007.

[5] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localisation*. Cambridge, MA: MIT, 1997.

[6] R. Martin, U. Heute, and C. Antweiler, *Advances in Digital Speech Transmission*. Hoboken, NJ: Wiley, 2008.

[7] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, pp. 276–280, Mar. 1986.

[8] C. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

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

[10] A. Lombard, Y. Zheng, H. Buchner, and W. Kellermann, "TDOA estimation for multiple sound sources in noisy and reverberant environments using broadband independent component analysis," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 19, no. 6, pp. 1490–1503, Aug. 2011.

[11] R. O. Duda and W. L. Martens, "Range dependence of the response of a spherical head model," *J. Acoust. Soc. Amer.*, vol. 104, no. 5, pp. 3048–3058, 1998.

[12] D. L. Wang and G. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*. New York: IEEE Press/Wiley-Interscience, 2007.

[13] T. May, S. van de Par, and A. Kohlrausch, "A probabilistic model for robust localization based on a binaural auditory front-end," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 19, no. 1, pp. 1–13, Jan. 2011.

[14] J. Scheuing and B. Yang, "Disambiguation of TDOA estimation for multiple sources in reverberant environments," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 16, no. 8, pp. 1479–1489, 2008.

[15] G. W. Elko, "Microphone array systems for hands-free telecommunication," *Speech Commun.*, vol. 20, no. 3–4, pp. 229–240, 1996.

[16] J. M. Kates and M. R. Weiss, "A comparison of hearing-aid array-processing techniques," *J. Acoust. Soc. Amer.*, vol. 99, no. 5, pp. 3138–3148, May 1996.

[17] S. Doclo and M. Moonen, "Superdirective beamforming robust against microphone mismatch," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 15, no. 2, pp. 617–631, Feb. 2007.

[18] E. Mabande, A. Schad, and W. Kellermann, "Design of robust superdirective beamformers as a convex optimization problem," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Taipei, Taiwan, Apr. 2009, pp. 77–80.

[19] T. Lotter and P. Vary, "Dual-channel speech enhancement by superdirective beamforming," *EURASIP J. Appl. Signal Processing*, vol. 2006, no. 1, pp. 175–175, Jan. 2006.

[20] B. V. Veen and K. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–24, 1988.

[21] J. Capon, "High resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 7, pp. 1408–1418, Aug. 1969.

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

[23] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Trans. Signal Processing*, vol. 47, no. 10, pp. 2677–2684, Oct. 1999.

[24] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and non-stationarity with applications to speech," *IEEE Trans. Signal Processing*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[25] A. Krueger, E. Warsitz, and R. Haeb-Umbach, "Speech enhancement with a GSC-like structure employing eigenvector-based transfer function ratios estimation," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 19, no. 1, pp. 206–218, Jan. 2011.

[26] H. Cox, R. M. Zeskind, and M. M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 35, no. 10, pp. 1365–1376, Oct. 1987.

[27] A. Spriet, M. Moonen, and J. Wouters, "Spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction," *Signal Process.*, vol. 84, no. 12, pp. 2367–2387, Dec. 2004.

[28] F. L. Luo, J. Y. Yang, C. Pavlovic, and A. Nehorai, "Adaptive null-forming scheme in digital hearing aids," *IEEE Trans. Signal Processing*, vol. 50, no. 7, pp. 1583–1590, 2002.

[29] J. B. Maj, L. Royackers, M. Moonen, and J. Wouters, "Comparison of adaptive noise reduction algorithms in dual microphone hearing aids," *Speech Commun.*, vol. 48, no. 8, pp. 957–960, Aug. 2006.

[30] B. Cornelis, M. Moonen, and J. Wouters, "Speech intelligibility improvements with hearing aids using bilateral and binaural adaptive multichannel Wiener filtering based noise reduction," *J. Acoust. Soc. Amer.*, vol. 131, no. 6, pp. 4743–4755, June 2012.

[31] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, "Frequency-domain criterion for speech distortion weighted multichannel Wiener filter for robust noise reduction," *Speech Commun.*, vol. 49, no. 7–8, pp. 636–656, July–Aug. 2007.

[32] L. W. Brooks and I. S. Reed, "Equivalence of the likelihood ratio processor, the maximum signal-to-noise ratio filter, and the Wiener filter," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 8, no. 5, pp. 690–692, 1972.

[33] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 3, pp. 320–327, May 2000.

[34] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 2, pp. 109–116, Mar. 2003.

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

[36] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, nos. 1–3, pp. 21–34, Nov. 1998.

[37] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. Speech Audio Processing*, vol. 12, no. 5, pp. 530–538, Sept. 2004.

[38] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source separation," in *Proc. Independent Component Analysis and Signal Separation*, Dec. 2001, pp. 722–727.

[39] 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 J. Appl. Signal Process.*, vol. 2003, no. 11, pp. 1157–1166, Nov. 2003.

[40] Y. Zheng, K. Reindl, and W. Kellermann, "Analysis of dual-channel ICA-based blocking matrix for improved noise estimation," *EURASIP J. Appl. Signal Processing*, vol. 26, 2014. Doi:10.1186/1687-6180-2014-26. [Online]. Available: http://asp.eurasipjournals.com/content/2014/1/26

[41] S. Nordholm, I. Claesson, and M. Dahl, "Adaptive microphone array employing calibration signals: An analytical evaluation," *IEEE Trans. Speech Audio Processing*, vol. 7, no. 3, pp. 241–252, May 1999.

[42] T. Gerkmann, C. Breithaupt, and R. Martin, "Improved a posteriori speech presence probability estimation based on a likelihood ratio with fixed priors," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 16, no. 5, pp. 910–919, July 2008.

[43] K. Reindl, S. Markovich-Golan, H. Barfuss, S. Gannot, and W. Kellermann, "Geometrically constrained TRINICON-based relative transfer function estimation in underdetermined scenarios," in *Proc. IEEE Workshop Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2013.

[44] T. Van den Bogaert, S. Doclo, J. Wouters, and M. Moonen, "The effect of multimicrophone noise reduction systems on sound source localization by users of binaural hearing aids," *J. Acoust. Soc. Amer.*, vol. 124, no. 1, pp. 484–497, July 2008.

[45] G. Grimm, V. Hohmann, and B. Kollmeier, "Increase and subjective evaluation of feedback stability in hearing aids by a binaural coherence-based noise reduction scheme," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 17, no. 7, pp. 1408–1419, Sept. 2009.

[46] K. Reindl, Y. Zheng, A. Schwarz, S. Meier, R. Maas, A. Sehr, and W. Kellermann, "A stereophonic acoustic signal extraction scheme for noisy and reverberant environments," *Comput. Speech Lang. (CSL)*, vol. 27, no. 3, pp. 726–745, May 2012.

[47] D. Welker, J. Greenberg, J. Desloge, and P. Zurek, "Microphone-array hearing aids with binaural output–Part II: A two-microphone adaptive system," *IEEE Trans. Speech Audio Processing*, vol. 5, no. 6, pp. 543–551, Nov. 1997.

[48] B. Cornelis, S. Doclo, T. Van den Bogaert, J. Wouters, and M. Moonen, "Theoretical analysis of binaural multi-microphone noise reduction techniques," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 18, no. 2, pp. 342–355, Feb. 2010.

[49] E. Hadad, S. Gannot, and S. Doclo, "Binaural linearly constrained minimum variance beamformer for hearing aid applications," in *Proc. Int. Workshop on Acoustic Signal Enhancement (IWAENC)*, Aachen, Germany, Sept. 2012, pp. 117–120.

[50] D. Marquardt, V. Hohmann, and S. Doclo, "Perceptually motivated coherence preservation in multi-channel Wiener filtering based noise reduction for binaural hearing aids," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Florence, Italy, May 2014, pp. 3688–3692.