Thiemann *et al. EURASIP Journal on Advances in Signal Processing* (2016) 2016:12
DOI 10.1186/s13634-016-0314-6

EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

RESEARCH Open Access

Speech enhancement for multimicrophone binaural hearing aids aiming to preserve the spatial auditory scene

Joachim Thiemann[1*], Menno Müller[1,2], Daniel Marquardt[1], Simon Doclo[1] and Steven van de Par[1]

Abstract

Modern binaural hearing aids utilize multimicrophone speech enhancement algorithms to enhance signals in terms of signal-to-noise ratio, but they may distort the interaural cues that allow the user to localize sources, in particular, suppressed interfering sources or background noise. In this paper, we present a novel algorithm that enhances the target signal while aiming to maintain the correct spatial rendering of both the target signal as well as the background noise. We use a bimodal approach, where a signal-to-noise ratio (SNR) estimator controls a binary decision mask, switching between the output signals of a binaural minimum variance distortionless response (MVDR) beamformer and scaled reference microphone signals. We show that the proposed selective binaural beamformer (SBB) can enhance the target signal while maintaining the overall spatial rendering of the acoustic scene.

Keywords: Hearing aids, Binaural hearing aids, Bilateral hearing aids, Binaural MVDR

1 Introduction

Situations where we are exposed to a number of sound sources reaching our ears simultaneously are part of our everyday life. In these situations, the human auditory system is able to focus intentionally on a single sound source while suppressing other interfering sound sources. This process is referred to as auditory scene analysis (ASA) [1]. One example of ASA is the so-called cocktail party effect [2, 3], which describes the human ability to follow a conversation with a single target speaker while other interfering speakers are active. The cocktail party effect is a major area of hearing research, and many important factors that are part of human ASA have been identified. One of these factors is the spatial separation of sound sources [3–5] which leads to a spatial release from masking (SRM). SRM is the increased ability to hear signals in noise if the signal and noise have different perceived directions.

Compared to normal-hearing persons, hearing-impaired persons find it more difficult to handle cocktail party situations. This can be explained by the fact that hearing-impaired persons have a higher speech reception threshold (SRT) for speech in noise and do not benefit as much from SRM as normal-hearing persons [3–6]. Hence, in hearing aid algorithms, it is important to increase the signal-to-noise ratio (SNR) of the desired sound source in order to improve speech intelligibility.

However, any modification of the signals presented to the ears has the potential to distort the cues the ear uses to perceive the direction of sounds (e.g., interaural level difference (ILD) and interaural time difference (ITD)) [7, 8]. This information should be preserved not only to maintain the benefit of SRM but also to allow the person to be aware of the spatial composition of his surroundings.

Miniaturization has allowed modern hearing aids to utilize multiple microphones in compact devices, allowing the use of multimicrophone signal enhancement algorithms [9–12]. Methods such as fixed and adaptive beamforming and multichannel Wiener filtering are generally capable of better noise suppression and lower speech distortion than single-channel methods [9, 11]. In the case of bilateral hearing loss, a hearing aid is required at both ears. Both hearing aids may work independently, but it is advantageous to link them together, treating all microphone channels as a unified array. This increases the array gain, but results in the signals presented to the ears being

*Correspondence: joachim.thiemann@uni-oldenburg.de
[1] University of Oldenburg, Cluster of Excellence "Hearing4All", Ammerländer Heerstr. 114-118, 26129 Oldenburg, Germany
Full list of author information is available at the end of the article

generated from the same set of input signals making it necessary to take measures to control the interaural cues.

Presently, methods that perform array processing while aiming to preserve interaural cues can be divided into two general categories. The first category of methods uses a (real-valued) spectro-temporal gain where the same gain is applied to two microphone signals (one on each hearing aid) [13–16]. By applying the same gain to both sides, these methods preserve all differences (both level and temporal) between the reference microphone signals but may suffer in noise reduction performance since they are effectively single-channel noise reduction methods, i.e., the signal at the left ear is generated by applying a spectro-temporal gain to the left reference microphone signal while the signal at the right ear is generated by applying the same spectro-temporal gain to the right microphone signal. The second category uses a more general approach, combining spectral and spatial filtering [9, 17, 18] based on a cost function which allows for controlling the amount of interaural cue distortion. Although typically a larger noise reduction can be achieved, there is always a trade-off between noise reduction performance and interaural cue preservation for interfering sources and the background noise.

In this paper, we will assume an acoustic scenario with a single target source, which is assumed to be a localized source, and background noise, which is assumed to be isotropic. For binaural signal enhancement, we suggest a bimodal processing paradigm: (a) where the target is dominant, use signal enhancement that preserves the interaural cues of the target; (b) where the background noise is dominant, pass the acoustic signal to the output unmodified save by attenuation. We realize this bimodal processing by using a binary decision mask in the spectro-temporal plane, i.e., within each frequency band and time frame, the signals presented to the ears are either the enhanced (binaural) target signal or the attenuated background noise (taken from two reference microphones, located near the respective ear). The challenge is to create an accurate SNR-dependent binary decision mask, since selecting the enhanced target signal for too many time-frequency bins will destroy the spatial cues of the background noise, while selecting the attenuated background for too many time-frequency bins will produce distortions for the target signal.

The proposed enhancement algorithm is implemented using the well-known binaural minimum variance distortionless response (MVDR) beamformer [9, 11], coupled with an SNR estimator based on the assumption that the background noise is isotropic [19]. In subjective evaluations using measurements of the speech reception threshold (SRT), we find that the overall enhancement performance is equivalent to other cue-preserving binaural algorithms such as MVDR with partial noise estimation. However, when subjects compared the spatial rendering of the overall acoustic scene, we find that, in situations where the beamformer target is off the center direction, our proposed algorithm preserves the spatial image better than the algorithms to which it was compared.

2 Background

In this paper, we consider a binaural hearing aid with a small number of microphones close to each ear, such as the device depicted in Fig. 1. The total number of microphones is denoted by M, and we assume that the hearing aids on both ears are linked. For each hearing aid, we consider one of the microphones to be the reference microphone, i.e., disregarding equalization effects the receiver signal would optimally be equal to the reference microphone signal in the absence of noise.

Working in the short-time Fourier transform (STFT) domain with f and n denoting the frequency and time indices, respectively, the M-dimensional microphone signal is given by $\mathbf{x}(f,n) = \left[x_1(f,n)\, x_2(f,n) \ldots x_M(f,n)\right]^T$. Assuming an acoustic scenario with a single localized target source $s(f,n)$, the microphone signal can be written as

$$\mathbf{x}(f,n) = \mathbf{d}_T(f)s(f,n) + \mathbf{v}(f,n), \tag{1}$$

where $\mathbf{d}_T(f)$ denotes the transfer function of the direct path between the target source to the microphones (including head shadow effects), which is assumed to be known. The overall noise component $\mathbf{v}(f,n)$ contains a mixture of ambient noise, interfering sources, and the (early and late) reverberation of the target source. The overall noise component is assumed to be uncorrelated to the target signal and spatially isotropic.[1] Since we treat all frequency bands independently, we omit the frequency index f in the following discussion without loss of generality.

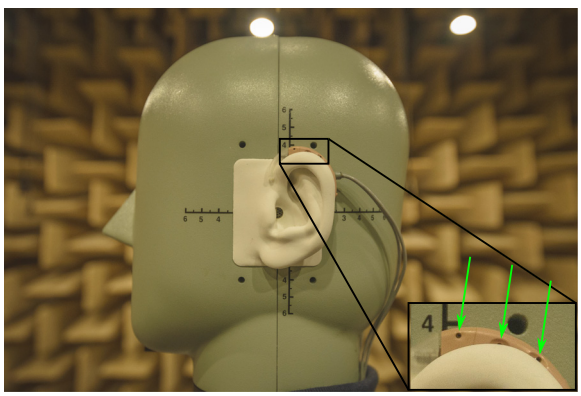

Fig. 1 Left side of a binaural multimicrophone hearing aid. In the cutout, the *arrows* indicate the location of the microphones. The opposite side has a symmetrical arrangement of microphones

Using the assumptions, the covariance matrix of $\mathbf{x}(n)$ can be written as

$$\begin{aligned}\mathbf{\Phi}_{\mathbf{x}}(n) &= E\{\mathbf{x}(n)\mathbf{x}^H(n)\} && (2)\\ &= E\{s^2(n)\}\mathbf{d}_T\mathbf{d}_T^H + \mathbf{\Phi}_v(n) && (3)\\ &\approx \phi_s(n)\mathbf{d}_T\mathbf{d}_T^H + \phi_v(n)\mathbf{\Gamma}_{\text{iso}}, && (4)\end{aligned}$$

where $E\{\cdot\}$ denotes the expected value operator, $\phi_s(n)$ is the power spectral density (PSD) of the target signal, $\phi_v(n)$ the PSD of the overall noise, $\mathbf{\Phi}_v$ the $M \times M$-dimensional noise covariance matrix, and $\mathbf{\Gamma}_{\text{iso}}$ the normalized covariance matrix of the spatially isotropic (diffuse) noise field [20]. In particular, we assume cylindrically isotropic noise, i.e., sound coming from all directions in the horizontal plane with equal probability.

2.1 The binaural MVDR beamformer

The core of our proposed algorithm is the well-known MVDR beamformer [21].[2] For the MVDR beamformer, the filter coefficient vector is equal to

$$\mathbf{w}_{\text{MVDR}}(n) = \frac{\mathbf{\Phi}_v^{-1}(n)\mathbf{d}_s(n)}{\mathbf{d}_s^H(n)\mathbf{\Phi}_v^{-1}(n)\mathbf{d}_s(n)}, \quad (5)$$

with $\mathbf{d}_s(n)$ the M-dimensional steering vector. For the steering vector, either the transfer function of the complete room impulse response (RIR) (including reverberation), the direct path of the RIR (corresponding to the free-field head-related transfer function (HRTF)) or the so-called relative transfer function can be used [11, 22]. In the proposed algorithm, we will use the free-field HRTF as the steering vector, assuming that the direction of the target source is known and time-invariant, i.e., $\mathbf{d}_s(n) = \mathbf{d}_T$. Assuming a spatially isotropic noise field, i.e., $\mathbf{\Phi}_v(n) = \phi_v(n)\mathbf{\Gamma}_{\text{iso}}$, the MVDR beamformer in (5) reduces to a fixed (superdirective) beamformer [20], i.e.,

$$\mathbf{w}_{\text{MVDR}} = \frac{\mathbf{\Gamma}_{\text{iso}}^{-1}\mathbf{d}_T}{\mathbf{d}_T^H\mathbf{\Gamma}_{\text{iso}}^{-1}\mathbf{d}_T}. \quad (6)$$

In the context of binaural hearing aids, the MVDR beamformer can be extended to provide a binaural output signal by defining two steering vectors that are normalized w.r.t. different reference microphones [9, 17], for which we have chosen the microphones closest to the left and right ear canals. Thus, given the steering vector $\mathbf{d}_T$ and denoting channels 1 and 2 as the left and right reference microphones, respectively, we define

$$\mathbf{d}_L = \frac{1}{d_1}\mathbf{d}_T \qquad \text{and} \qquad \mathbf{d}_R = \frac{1}{d_2}\mathbf{d}_T, \quad (7)$$

where d_1 and d_2 denote the first and second elements of $\mathbf{d}_T$. The left and right filter coefficient vectors are equal to

$$\mathbf{w}_L = \frac{\mathbf{\Gamma}_{\text{iso}}^{-1}\mathbf{d}_L}{\mathbf{d}_L^H\mathbf{\Gamma}_{\text{iso}}^{-1}\mathbf{d}_L} = d_1^*\mathbf{w}_{\text{MVDR}} \quad (8)$$

and similarly, $\mathbf{w}_R = d_2^*\mathbf{w}_{\text{MVDR}}$, such that the beamformer outputs $y_{\text{bf}L}$ and $y_{\text{bf}R}$ are obtained as

$$\begin{aligned}y_{\text{bf}L} &= \mathbf{w}_L^H\mathbf{x}(n) && (9)\\ &= d_1\left(\mathbf{w}_{\text{MVDR}}^H\mathbf{d}_T s(n) + \mathbf{w}_{\text{MVDR}}^H\mathbf{v}(n)\right) && (10)\\ &= d_1\left(s(n) + \mathbf{w}_{\text{MVDR}}^H\mathbf{v}(n)\right), && (11)\\ y_{\text{bf}R} &= \mathbf{w}_R^H\mathbf{x}(n) = d_2\left(s(n) + \mathbf{w}_{\text{MVDR}}^H\mathbf{v}(n)\right), && (12)\end{aligned}$$

where we use the bf subscript to indicate the beamformer output. Since the resulting left and right output signals only differ by a linear time-invariant filter (d_1 and d_2), this signal will be perceived as a point source in the direction of the target source, preserving the interaural cues of the target source but typically destroying the cues of the background noise [9, 17].

3 Selective binaural beamformer

The main idea behind the proposed algorithm is to preserve the spatial impression of the overall acoustic scene, i.e., both for the target sources and for the overall background noise. As described above, the binaural MVDR beamformer will cause the entire signal to be perceived as a point source in the direction of the target source. This is clearly undesired since not only would the hearing aid user lose the acoustic impression of the space in which he finds himself, but more importantly, there are many situations in which it is crucial for the hearing aid user to be able to localize a sound that is in the background (e.g., a new speaker trying to gain the attention of the user or a warning signal from an approaching vehicle).

Designing a beamformer such that all interaural cues of the various sources are preserved is difficult. However, the binaural scene is available from the reference microphones, and we can use these signals for rendering the background noise. The signals presented to the user are constructed by switching, for each time-frequency (T-F) bin, between the enhanced signals and an attenuated version of the unprocessed signals from the reference microphones. We term this approach the selective binaural beamformer (SBB).

To decide whether to present the enhanced beamformer output signals or the signals from the reference microphones, we compare the PSD of the target signal $\phi_s(n)$ to the PSD of the overall noise (from all directions) $\phi_v(n)$, equivalent to computing the input SNR $\phi_s(n)/\phi_v(n)$ and comparing it to a threshold value of 0 dB. In T-F bins where $\phi_s(n)$ exceeds $\phi_v(n)$, the enhanced signals are used; otherwise, the signals from the reference microphones are passed through with some attenuation, as shown schematically in Fig. 2. For example, if the target signal is a speech signal at low SNR, most bins in the T-F plane will be classified as background noise rather than target signal. Since the background noise is only modified by (possibly frequency-dependent) attenuation, its spatial character is

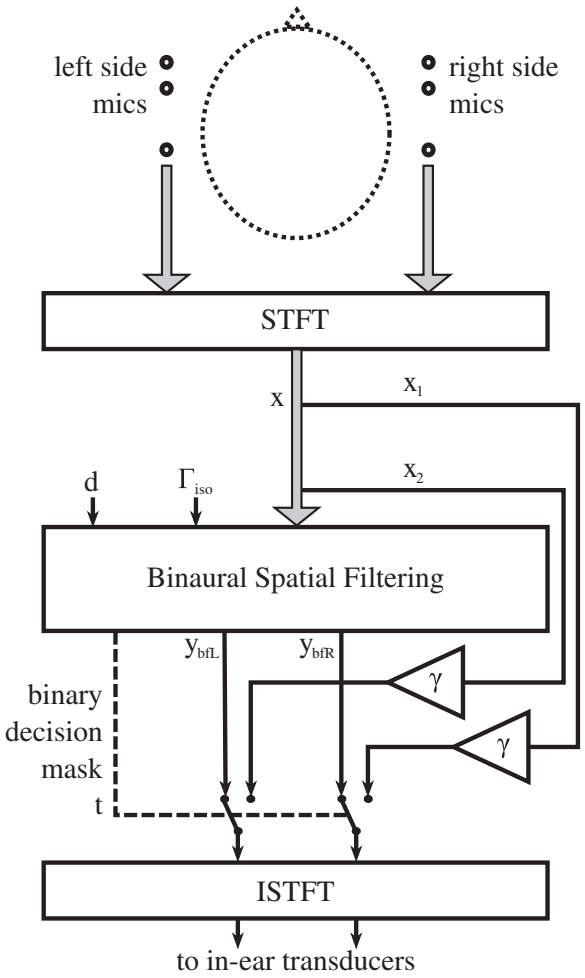

Fig. 2 Schematic of beamforming with binary decision mask. The multichannel input signal is converted into the time-frequency (T-F) domain and then passed to the binaural spatial filtering algorithm. The binary decision mask, based on the SNR estimator, then either passes the output signals of the binaural spatial filter or the scaled reference microphone signals to the in-ear transducers

preserved. Those portions of the T-F plane dominated by the target signal, on the other hand, will be perceived from the target direction.

There are several factors that determine the quality of the enhanced binaural signals, with two extreme cases that can be viewed as quality limits. First, in the case where no T-F bins are classified as being from the target source, the spatial scene is perfectly preserved; however, no enhancement takes place. On the other hand, if the complete T-F plane is classified as being from the target source, the entire auditory scene collapses to the target direction. The desired compromise therefore heavily depends on the accuracy of the binary decision mask.

If perfect knowledge of the target signal and the background noise are available, we can construct a so-called ideal binary decision mask (IBDM), differentiating all T-F bins between target ($\phi_s(n) > \phi_v(n)$) and noise ($\phi_v(n) > \phi_s(n)$). Deviations from this decision mask due to estimation errors of the input SNR will occur for T-F bins where the target signal is falsely classified as noise or where the noise is classified as being part of the target signal. In the former case, the target signal will be attenuated just as much as the background noise. In the latter case, the background noise will loose its spatial characteristics. It might be assumed from this observation that the overall quality of the enhanced signals is degraded more by underestimating the input SNR; however, overestimating the input SNR results in a decision mask containing spurious misclassifications which introduces disturbing artifacts.

3.1 SNR estimation

In the context of beamforming algorithms, estimating the SNR is a common problem for the design of single-channel postfilters [23–26]. Based on the signal model in (1), we will use the maximum likelihood estimator (MLE) proposed in [19, 26], which requires an estimate of the covariance matrix of the input signal, knowledge about the steering vector of the target source, and the scaled noise covariance matrix. The estimates of the spectral variance of the noise and the target signal are given by

$$\hat{\phi}_v(n) = \frac{1}{M-1}\mathrm{tr}\left[\left(\mathbf{I}_M - \mathbf{d}_T\mathbf{w}_{\mathrm{MVDR}}^H\right)\hat{\Phi}_{\mathbf{x}}(n)\mathbf{\Gamma}_{\mathrm{iso}}^{-1}\right] \quad (13)$$

$$\hat{\phi}_s(n) = \mathbf{w}_{\mathrm{MVDR}}^H\left(\hat{\Phi}_{\mathbf{x}}(n) - \hat{\phi}_v(n)\mathbf{\Gamma}_{\mathrm{iso}}^{-1}\right)\mathbf{w}_{\mathrm{MVDR}}, \quad (14)$$

where $\hat{\Phi}_{\mathbf{x}}(n)$ is an estimate of the covariance matrix of the input signal, $\mathbf{I}_M$ is the $M \times M$ identity matrix, and $\mathrm{tr}\{\cdot\}$ denotes the matrix trace. While in [19] these two estimates were used to compute the postfilter, in our algorithm, we will use the estimated input SNR ($\hat{\phi}_s(n)/\hat{\phi}_v(n)$) in each T-F bin for the binary decision mask.

3.2 Binary decision mask

The binary decision mask (BDM) generated using the SNR estimate described above can be expressed as

$$t(n) = \begin{cases} 1, & \frac{\hat{\phi}_s(n)}{\hat{\phi}_v(n)} > 1, \\ 0, & \text{otherwise.} \end{cases} \quad (15)$$

In other words, the decision mask is equal to 0 for all T-F bins where the local input SNR estimate is smaller than the threshold value of 0 dB, which is the minimum SNR required for listeners to detect usable glimpses from the target speech signal that will aid intelligibility [27]. A higher threshold would potentially distort the speech signal [28], whereas a lower threshold impacts the perception of the spatial characteristics of the background noise. Note that in practice, we found that varying the threshold by ±3 dB does not have a noticeable effect.

The BDM is directly used for constructing the output signals, with the output signal presented to the left ear equal to

$$y_{\mathrm{SBB},L}(n) = \begin{cases} y_{\mathrm{bf}L}(n), & t(n) = 1, \\ \gamma x_{m_L}(n), & t(n) = 0, \end{cases} \tag{16}$$

where $\gamma(f)$ is a possibly frequency-dependent scaling factor and $x_{m_L}(n)$ is the left reference microphone signal. The output $y_{\mathrm{SBB},R}$ is computed in a similar manner using $y_{\mathrm{bf}R}(n)$ and $x_{m_R}(n)$. The scaling factor $\gamma(f)$ controls the amount of noise attenuation affecting the amount of artifacts produced in the output.

In order to avoid large gain variations for the residual noise when switching between the beamformer outputs and the reference microphone signals, we need to compensate for the array gain of the binaural MVDR beamformer, which for the left and the right side, respectively, is equal to

$$G_L = (\mathbf{w}_L^H \mathbf{\Gamma}_{\mathrm{iso}} \mathbf{w}_L)^{-1} \quad \text{and} \tag{17}$$

$$G_R = (\mathbf{w}_R^H \mathbf{\Gamma}_{\mathrm{iso}} \mathbf{w}_R)^{-1}. \tag{18}$$

In order to preserve the ILDs of the background noise, we have used an equal scaling factor on both sides, where we have used the "better ear" side since the user's attention will be towards this side, i.e.,

$$\gamma = \begin{cases} G_L, & \text{target is on left side,} \\ G_R, & \text{target is on right side.} \end{cases} \tag{19}$$

Since this scaling factor is frequency-dependent, it will result in a spectral coloration of the background noise. Alternatively, we will therefore also consider a constant (frequency-independent) scaling factor γ_c by averaging γ over all frequencies [29].

3.3 MVDR-N

Preserving the interaural cues of the background noise is a problem that has been addressed by other binaural beamforming algorithms, and it is interesting to compare the SBB in particular to the binaural MVDR beamformer with partial noise estimation (MVDR-N). The MVDR-N is an extension of the binaural MVDR beamformer, mixing the output signals of the binaural MVDR with the reference microphone signals in order to provide a trade-off between noise reduction and preservation of the interaural cues of the noise component [17, 30, 31], i.e.,

$$y_{\text{MVDR-N,L}} = (1-\eta) y_{\mathrm{bf}L} + \eta x_1, \tag{20}$$

$$y_{\text{MVDR-N,R}} = (1-\eta) y_{\mathrm{bf}R} + \eta x_2. \tag{21}$$

Hence, more generally, both the SBB and the MVDR-N can be described as mixing the binaural MVDR beamformer output signals and the reference microphone signals using

$$y_{\mathrm{gen,L}} = \alpha y_{\mathrm{bf}L} + \beta x_1, \tag{22}$$

$$y_{\mathrm{gen,R}} = \alpha y_{\mathrm{bf}R} + \beta x_2, \tag{23}$$

with α and β shown in Table 1. The main differences between the SBB and the MVDR-N lie in the fact that for the SBB, the parameters α and β depend on the estimated SNR and are hence signal- and frequency-dependent.

4 Evaluation

Using the binaural database presented in [32], we simulate noisy and reverberant microphone signals by convolving clean speech samples from the Oldenburger Satztest (OLSA) database [33–35], with room impulse responses measured on an artificial head and by adding noise recorded by the same setup in the same room. The used sampling rate is 16 kHz. The room is a university cafeteria ($T_{60} \approx 1250$ ms), and the noise is a typical ambient noise for this environment, i.e., a combination of many speakers in different positions plus incidental noises from cutlery, chairs, etc. Both hearing aids consist of two microphones (thus $M = 4$ microphones in total), where the distance between the two microphones is 15.6 mm.

In the evaluation, we use two different spatial setups. In the first setup, the target speaker is in front of the hearing aid user. In the second setup, the target speaker is on the left side of the user (30° azimuth angle). For the steering vector $\mathbf{d}_T$, we have used the anechoic HRTFs, also from [32]. This corresponds to the early reflections being treated as interfering noise but yields a realistic implementation since the real acoustic path is generally not known.

The algorithm is implemented using frame-based short-time Fourier transform (STFT) processing with a frame size of 320 samples (zero-padded to 512 samples for the FFT) corresponding to 20 ms, and a frame advance of 160 samples (corresponding to 10 ms).

We estimate the noise coherence matrix $\mathbf{\Gamma}_{\mathrm{iso}}$ assuming a cylindrically isotropic noise field by averaging the

Table 1 Comparison of mixing factors for binaural beamforming algorithms

	MVDR	MVDR-N	SBB
α	1	$1-\eta$	1 $\hat{\phi}_s(n)/\hat{\phi}_v(n) > 1$ 0 otherwise
β	0	η	0 $\hat{\phi}_s(n)/\hat{\phi}_v(n) > 1$ γ otherwise

anechoic acoustic transfer functions (ATF) along the horizontal plane, as

$$\mathbf{\Gamma}_{\text{iso}}^{[m,n]} = \frac{\sum_{k=1}^{K} D_m(k) D_n^*(k)}{\sqrt{\sum_{k=1}^{K} |D_m(k)|^2 \sum_{k=1}^{K} |D_n(k)|^2}}, \tag{24}$$

where $\mathbf{\Gamma}_{\text{iso}}^{[m,n]}$ is the (m,n)th element of the noise coherence matrix, K is the number of azimuth angles, and $D_m(k)$ is the M-dimensional anechoic ATF on microphone m from direction index k. We use 5° spacing ($K = 72$) using the HRTFs from [32].

For the SNR estimation (13, 14), the covariance matrix of the input signal was estimated by exponential averaging using

$$\hat{\Phi}_{\mathbf{x}}(n) = (1-\alpha)\mathbf{x}(n)\mathbf{x}^H(n) + \alpha\hat{\Phi}_{\mathbf{x}}(n-1), \tag{25}$$

where $\alpha = 0.7$, giving a time constant of 28 ms.

The proposed algorithm is tested in two variants: (a) with the frequency-dependent scaling factor $\gamma(f)$ in (16) (labeled SBB) and (b) with a frequency-independent scaling factor γ_c, computed by averaging γ over frequency (labeled SBB/CM, where the "CM" stands for "constant mixfactor"). In addition, to evaluate the impact of SNR estimation errors, we also used the ideal binary decision mask (IBDM) from oracle information, i.e., using the unmixed target and noise signals. For the IBDM, we also considered a frequency-dependent and a frequency-independent scaling factor, labeled as IBDM and IBDM/CM.

As reference algorithms, we used three variants of the MVDR beamformer: (a) a bilateral MVDR beamformer where each hearing aid operates independently of the other (i.e., two separate two-microphone beamformers on each hearing aid), labeled "BIL", (b) a non-spatial cue preserving binaural MVDR beamformer (i.e., using $y_{\text{bf}L}$ and $y_{\text{bf}R}$), and (c) the binaural MVDR beamformer with partial noise estimation as described in Section 3.3, labeled "MVDR-N," where we have used the frequency-independent mixing factor $\eta = 0.2$ as proposed in [36].

4.1 Instrumental evaluation

The noise reduction performance of the proposed algorithm is evaluated using the intelligibility weighted SNR (iSNR) [37], taking the maximum value of the two ear channels to simulate the better ear effect. Speech quality is evaluated using the perceptual evaluation of speech quality (PESQ) measure [38], where the clean speech signal at the reference microphones is used as reference.

Our test signals are male and female speech rendered using binaural room impulse responses and mixed at iSNR of −5.7 dB for the target source in front and 0 dB for the target at the side.

Tables 2 and 3 present the instrumental performance measures for the considered algorithms, showing the difference in iSNR to the reference microphone signal (ΔiSNR), the PESQ score (in terms of mean opinion score (MOS)), the ILD error of the target signal ($\Delta\text{ILD}_{\text{tg}}$, in dB), the ITD error of the target signal ($\Delta\text{ITD}_{\text{tg}}$, in μs), and the magnitude squared coherence (MSC) error of the background noise ($\Delta\text{MSC}_{\text{bg}}$). The ILD and the ITD are calculated based on the binaural model proposed in [39]. The MSC is defined as the square of the absolute value of the interaural coherence [40]. For the ILD, ITD, and MSC, the error is computed as the mean difference between the reference microphone signal cues, and the algorithm output signal cues, considering either only the target signal component (for $\Delta\text{ILD}_{\text{tg}}$ and $\Delta\text{ITD}_{\text{tg}}$) or the background noise component (for $\Delta\text{MSC}_{\text{bg}}$).

Table 2 Objective measures for target at center position

	ΔISNR (dB)	PESQ (MOS)	$\Delta\text{ILD}_{\text{tg}}$ (dB)	$\Delta\text{ITD}_{\text{tg}}$ (μs)	$\Delta\text{MSC}_{\text{bg}}$
BIL	3.73	1.74	1.50	1.59	0.06
MVDR	6.27	1.78	1.99	11.10	0.93
MVDR-N	5.25	1.75	2.09	7.31	0.66
IBDM	10.01	2.30	0.61	8.50	0.13
IBDM/CM	8.58	2.15	0.59	8.38	0.13
SBB	7.83	2.09	2.27	7.81	0.11
SBB/CM	6.69	2.00	2.04	7.60	0.11

The advantage provided by the binaural MVDR-based algorithms using four microphones compared to the bilateral MVDR beamformer is clear. In the case of the target source in front, the SBB algorithms show higher iSNR and PESQ scores than the binaural MVDR beamformer on which they are based. This can be explained by the fact that in the SBB output, the noise in the non-speech portions of the T-F plane is attenuated more strongly than in the MVDR output. This effect is not observed if the target source is to the side, since in the contralateral side, the binaural MVDR beamformer attenuates the noise more strongly than estimated by γ, which is based on the better ear side where the SNR gain is lower.

The importance of the SNR estimate can be observed from the iSNR scores for IBDM and IBDM/CM versus

Table 3 Objective measures for target at side position

	ΔISNR (dB)	PESQ (MOS)	$\Delta\text{ILD}_{\text{tg}}$ (dB)	$\Delta\text{ITD}_{\text{tg}}$ (μs)	$\Delta\text{MSC}_{\text{bg}}$
BIL	4.50	1.62	1.09	4.21	0.08
MVDR	10.87	1.90	1.74	13.56	0.91
MVDR-N	8.23	1.87	1.43	12.38	0.46
IBDM	7.46	1.87	0.79	17.86	0.03
IBDM/CM	5.73	1.76	0.94	18.57	0.02
SBB	5.79	1.86	1.54	17.26	0.05
SBB/CM	4.53	1.74	0.43	18.60	0.04

SBB and SBB/CM. Furthermore, it can be observed that using (the frequency-dependent) γ (IBDM and SBB) yields a higher performance than using the (frequency-independent) γ_c (IBDM-CM and SBB-CM). Also evident is the mild loss of performance (in terms of iSNR) of MVDR-N versus MVDR due to the mixing with the reference microphone signals [17, 40].

The ΔILD for the target source is observed to be small (always within 2.3 dB) for all algorithms. The ΔITD for the target source also indicates that we can expect all algorithms perform approximately equally well with respect to the binaural cues of the target. The bilateral MVDR beamformer configuration (BIL) appears to yield the best performance in terms of ΔITD, but this can be explained by the low iSNR gain this algorithm provides.

The ΔMSC shows the near complete removal of the spatial characteristics of the background noise for the binaural MVDR beamformer. The MVDR-N performs slightly better due to the partial noise estimation, but we clearly observe that the binary decision mask algorithms (IBDM, IBDM/CM, SBB, and SBB/CM) result in much lower MSC error, yielding a performance similar to the bilateral MVDR beamformer. In Section 4.2, we compare these instrumental measures with subjective evaluations.

4.1.1 *Influence of misestimating the steering vector*

Since the SBB requires an estimate of the direction of arrival (DOA) of the target source for both the underlying MVDR beamformer and to estimate the SNR, in this section, we evaluate the influence of misestimating the steering vector $\mathbf{d}_T$ on the iSNR. Figure 3 shows the ΔiSNR at the better ear for different assumed steering vectors $\mathbf{d}_T$ in the horizontal plane, for a target source in front (panel a) and a target source at 30° to the left (panel b). For both target source directions, the results for the SBB and the binaural MVDR are shown.

For small DOA estimation errors, it can be observed that the performance of the SBB is similar to the performance of the binaural MVDR beamformer. However, for large DOA estimation errors, the performance of the SBB tends to remain around $\Delta iSNR = 0$ dB, whereas the MVDR beamformer shows a significant iSNR decrease. This can be explained by the observation that for large DOA estimation errors, the SNR estimate will typically be below 0 dB, such that the binary decision mask is equal to 0, and the resulting SBB output signal is equal to the scaled reference microphone signal.

Interestingly, in our test scenario where the background noise contains an interfering speaker at 90°, the ΔiSNR for the SBB does drop below 0 dB if the assumed DOA is near to 90°. This can be explained by the observation that the SNR estimator uses the interfering speaker as target, such that the SBB attempts to enhance this interfering speaker, reducing the iSNR.

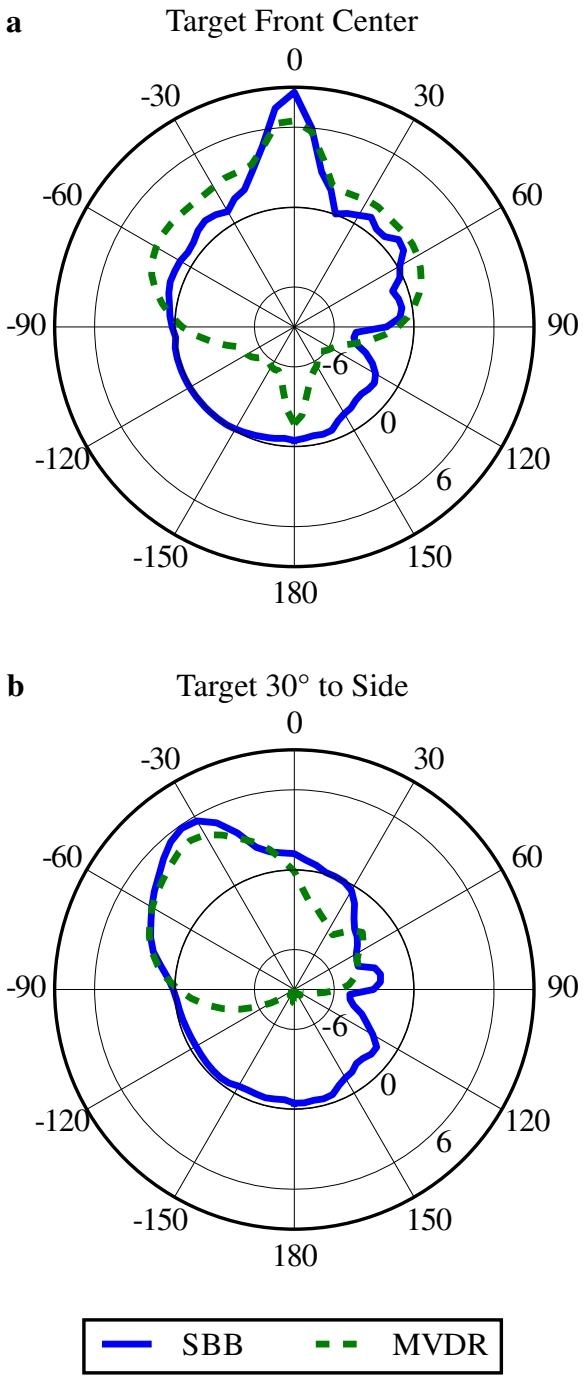

Fig. 3 ΔiSNR as a function of $\mathbf{d}_T$ for the target to the front (**a**) and the target at 30° to the left (**b**). The *solid blue line* shows the iSNR improvement for the SBB algorithm, the *dashed green line* for the binaural MVDR beamformer. The background noise includes an interfering speaker at 90°

4.2 Subjective evaluation

The subjective evaluation consists of two different tests, aiming to determine both the preservation of the

interaural cues of each algorithm as well as to measure speech intelligibility using the SRT. Eight self-reported normal-hearing subjects participated in the experiments, with ages ranging from 18 to 38 years. The processed signals were presented to the subjects via headphones (Sennheiser HD 650) connected to a USB Soundcard (RME Fireface UC) in a listening cabin at a level of 65 dB SPL. The measurements included a training phase during which the participants could listen to the signals without judging them.

4.2.1 Preservation of spatial rendering

To examine the preservation of the spatial rendering by the considered algorithms, the subjects participated in a test based on the multiple stimuli with hidden reference and anchor (MUSHRA) test [41]. The participants were asked to rate the generated binaural signals from the algorithms relative to the reference signal given by the unprocessed microphone signals. They were asked specifically to assess the spatial positioning of all sources in the acoustic scene, not just the target speaker but also the speech and non-speech elements of the background noise. For each test signal, the participants could rate the spatial rendering on a scale with the following labels: equal (100–80), almost equal (80–60), slightly different (60–40), different (40–20), and very different (20–0). Analogous to MUSHRA, we used the reference signal as test signal (hidden reference) and presented a low-quality anchor consisting of an averaged reference microphone signal presented to both ears. The same test signals as used for the objective evaluation were used, and each participant repeated the evaluation with a pause between sessions. No statistical difference was found in the evaluation scores of male or female target speakers, thus all four sets of responses of individual subjects were averaged. However, since one subject had difficulties with the evaluation, which was apparent by observing that the anchor was consistently not found (receiving a score greater than 50 in two instances), the scores from this subject were removed from the analysis.

Figure 4 shows the MUSHRA scores for the different algorithms, both for the target front center as well as for the target to the side. From the results, we observe that the SBB/CM algorithm performs slightly better than both the BIL and MVDR-N algorithms, especially when the target is to the side. As expected, MVDR performs badly, with MVDR-N showing considerably better scores. The bilateral arrangement (BIL) performs quite well, but when comparing the entire set of responses for both target positions, the distributions of subjective scores of SBB/CM versus BIL show a statistically significant difference with $p < 0.05$ using Welch's unequal variances t test. If we consider the scores for the target to the side only, we find $p < 0.01$. Comparing SBB/CM to MVDR-N, we also see

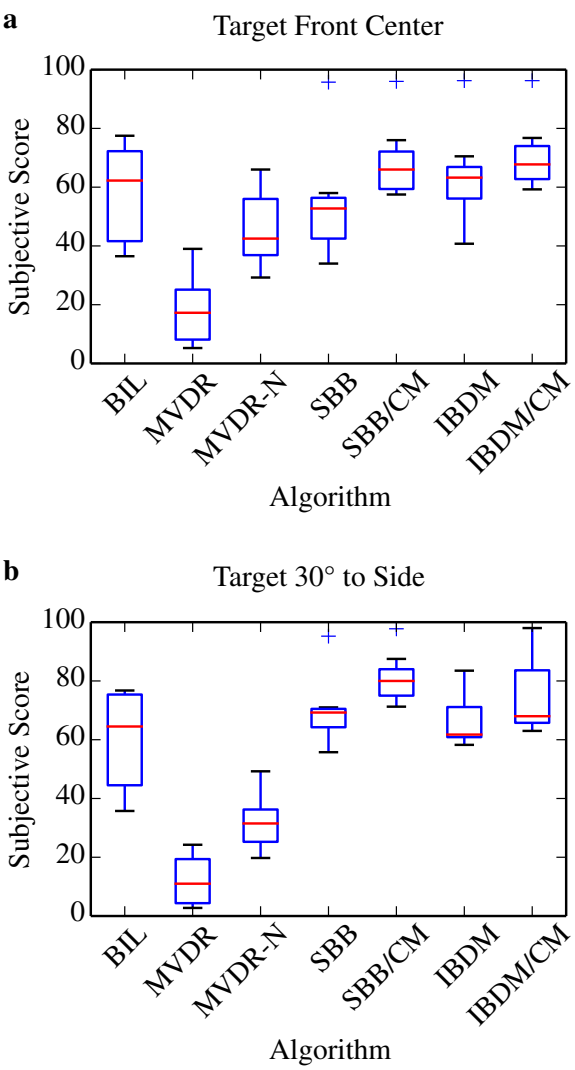

Fig. 4 Subjective evaluation of spatial rendering relative to the reference microphone signal, for the target in front (**a**) and for the target to the side (**b**). The *box* shows the extent of the middle quartiles (25 to 75 %) of the results with the *red center line* indicating the median. *Whiskers* show the data range up to 1.5 times the interquartile distance, and *crosses* show outliers

that the responses are statistically significantly different with $p < 0.01$ when evaluating the entire set of responses for both target responses. In the results, we also note that using a frequency-independent scale factor results in slightly better performance. Finally, we note that using an IBDM does not seem to translate into better performance compared to SBB, suggesting that the SNR estimator does not have a major influence in this regard. Comparing Fig. 4a to Table 2 and Fig. 4b to Table 3, we note that the subjective results appear to generally match the ΔMSC values, under the assumption that all algorithms render the target source at the correct spatial direction.

4.2.2 Speech reception threshold

The SRT is defined here as the SNR where 50 % of the words in a sentence are recognized by the listener [42]. It was measured for the considered beamforming algorithms and the unprocessed reference microphone signals using the OLSA [33–35] database and test procedure, for both target source positions, resulting in a total number of 16 runs. The OLSA test is a matrix test that consists of sentences with five words where each word can be one of ten alternatives. Each run consisted of a set of randomly chosen sentences from a list of 20 sentences. All participants performed the evaluation twice, with the SRTs averaged over the two sessions. The results for both target positions are shown in Fig. 5, where the SRT improvement is shown as the difference between the SRT of the unprocessed reference microphone signals and SRT of the processed signals, reducing the variability between subjects.

The results show that the SRT improvement of the SBB/CM is similar to the SRT improvement of the bilateral MVDR beamformer (using the complete set of responses, Welch's t test yields $p > 0.2$), but that the SBB/CM shows a significantly better preservation of the spatial cues. However, the SRT of the SBB/CM is worse by about 1 dB compared to the binaural MVDR and MVDR-N beamformer, although the difference is less when the target is to the side.

When the IBDM is used, the SRT is significantly lower than even the binaural MVDR. This can be explained both by the higher iSNR (cf. Tables 2 and 3) and the fact that the spectro-temporal activation of speech itself conveys information [43]. It also shows that the speech intelligibility improvement of SBB can be improved significantly if the accuracy of the SNR estimate can be improved.

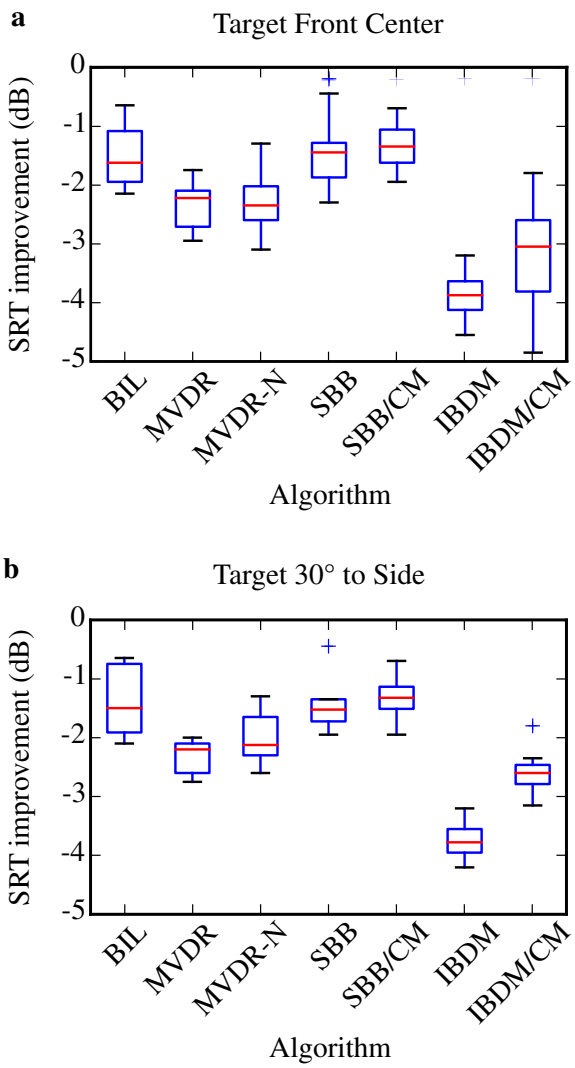

Fig. 5 SRT improvement relative to the reference microphone signal measured for the different algorithms, for the target in front (**a**) and for the target to the side (**b**). As in the previous plot, the *boxes* show the middle quartiles and the *red center line* the median. The SRT was measured as the SNR for 50 % word recognition

5 Conclusions

For an acoustic scene with a single target speaker and isotropic background noise, we have proposed a selective binaural beamformer, aiming to enhance the speech quality and intelligibility while preserving the spatial impression of the acoustic scene. The proposed selective binaural beamformer is based on a bimodal processing approach, where each time-frequency unit is considered to be either part of the target sound or the background noise. Based on an SNR estimator, a binary decision mask is constructed which decides if the output signals of a binaural MVDR beamformer (preserving the interaural cues of the target source) or the scaled reference microphone signals (preserving the spatial characteristics of the background noise) are used as the output signals.

Evaluating the results, we find that the SBB/CM algorithm provides a better preservation of the perceived spatial locations of sounds compared to other cue-preserving algorithms (such as the related MVDR-N and the bilateral MVDR beamformer). While there is some degradation of speech intelligibility compared to the binaural MVDR and MVDR-N, we achieve performance roughly on par with a bilateral MVDR beamformer. However, based on the results using the ideal binary decision mask, we expect that some of this degradation can be attributed to SNR estimation error.

Endnotes

[1] Although this assumption will not hold perfectly for the early reverberation of the target source and for the interfering sources, it is a commonly made assumption

for the derivation of superdirective beamformers [11, 20] and for dereverberation algorithms [19, 23].

[2]Other binaural noise reduction algorithms, such as the binaural multichannel Wiener filter [9], could also be used as the core algorithm.

Abbreviations

ASA: auditory scene analysis; ATF: acoustic transfer function; CM: constant mixfactor; BDM: binary decision mask; HRTF: head-related transfer function; IBDM: ideal binary decision mask; ILD: interaural level difference; ITD: interaural time difference; iSNR: intelligibility weighted signal-to-noise ratio; MLE: maximum likelihood estimate; MUSHRA: multiple stimulus with hidden reference and anchor; MVDR: minimum variance distortionless response; MVDR-N: MVDR with partial noise estimation; OLSA: Oldenburger Satztest; PESQ: perceptual evaluation of speech quality; PSD: power spectral density; RIR: room impulse response; SBB: selective binaural beamformer; SNR: signal-to-noise ratio; SRT: speech reception threshold; STFT: short-time Fourier transform; T-F: time-frequency.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This work was funded by the Deutsche Forschungsgemeinschaft EXC 1077, Cluster of Excellence "Hearing4All" (http://hearing4all.eu). We would like to thank the reviewers for their comments which helped to greatly improve this article. We would also like to thank the participants of the listening tests for their time and effort as well as our colleagues within the Cluster of Excellence "Hearing4All" for many insightful discussions.

Author details

[1]University of Oldenburg, Cluster of Excellence "Hearing4All", Ammerländer Heerstr. 114-118, 26129 Oldenburg, Germany. [2]Jade Hochschule, Ofener Str. 16/19, 26121 Oldenburg, Germany.

Received: 30 July 2015 Accepted: 20 January 2016
Published online: 02 February 2016

References

1. AS Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. (MIT press, Cambridge, 1994)
2. EC Cherry, Some experiments on the recognition of speech, with one and with two ears. J. Acoust. Soc. Am. **25**(5), 975–979 (1953)
3. AW Bronkhorst, The cocktail party phenomenon: a review of research on speech intelligibility in multiple-talker conditions. Acta Acust. united Ac. **86**(1), 117–128 (2000)
4. J Peissig, B Kollmeier, Directivity of binaural noise reduction in spatial multiple noise-source arrangements for normal and impaired listeners. J. Acoust. Soc. Am. **101**(3), 1660–1670 (1997). doi:10.1121/1.418150
5. ML Hawley, RY Litovsky, JF Culling, The benefit of binaural hearing in a cocktail party: effect of location and type of interferer. J. Acoust. Soc. Am. **115**(2), 833–843 (2004). doi:10.1121/1.1639908
6. R Beutelmann, T Brand, Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners. J. Acoust. Soc. Am. **120**(1), 331–342 (2006). doi:10.1121/1.2202888
7. J Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. (MIT Press, Cambridge, 1996)
8. T Van den Bogaert, S Doclo, J Wouters, M Moonen, The effect of multimicrophone noise reduction systems on sound source localization by users of binaural hearing aids. J. Acoust. Soc. Am. **124**(1), 484–497 (2008). doi:10.1121/1.2931962
9. S Doclo, S Gannot, M Moonen, A Spriet, in *Handbook on Array Processing and Sensor Networks*, ed. by S Haykin, KJR Liu. Chapter 9: acoustic beamforming for hearing aid applications (Wiley-IEEE Press, Hoboken, 2010), pp. 269–302
10. J Wouters, S Doclo, R Koning, T Francart, Sound processing for better coding of monaural and binaural cues in auditory prostheses. Proc. IEEE. **101**(9), 1986–1997 (2013). doi:10.1109/JPROC.2013.2257635
11. S Doclo, W Kellermann, S Makino, SE Nordholm, Multichannel signal enhancement algorithms for assisted listening devices: exploiting spatial diversity using multiple microphones. IEEE Signal Process. Mag. **32**(2), 18–30 (2015). doi:10.1109/MSP.2014.2366780
12. V Hamacher, U Kornagel, T Lotter, H Puder, *Binaural Signal Processing in Hearing Aids: Technologies and Algorithms*. (Wiley, New York, 2008), pp. 401–429. doi:10.1002/9780470727188.ch14
13. T Lotter, P Vary, Dual-channel speech enhancement by superdirective beamforming. EURASIP J. on Applied Sig. Proc. **2006**, 1–14 (2006). doi:10.1155/ASP/2006/63297
14. T Rohdenburg, V Hohmann, B Kollmeier, in *Proceedings of the Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. Robustness analysis of binaural hearing aid beamformer algorithms by means of objective perceptual quality measures (IEEE, New Paltz, 2007), pp. 315–318. doi:10.1109/ASPAA.2007.4393016
15. K Reindl, Y Zheng, W Kellermann, in *Proceedings of the European Signal Processing Conference (EUSIPCO)*. Analysis of two generic Wiener filtering concepts for binaural speech enhancement in hearing aids, (Aalborg, 2010), pp. 989–993
16. JI Marin-Hurtado, DN Parikh, DV Anderson, Perceptually inspired noise-reduction method for binaural hearing aids. IEEE Trans. Audio, Speech, Language Process. **20**(4), 1372–1382 (2012). doi:10.1109/TASL.2011.2179295
17. B Cornelis, S Doclo, T Van dan Bogaert, M Moonen, J Wouters, Theoretical analysis of binaural multimicrophone noise reduction techniques. IEEE Trans. Audio, Speech, Language Process. **18**(2), 342–355 (2010). doi:10.1109/TASL.2009.2028374
18. D Marquardt, V Hohmann, S Doclo, in *Proceedings of the International Conference on Acoustics , Speech and Signal Processing (ICASSP)*. Coherence preservation in multi-channel Wiener filtering based noise reduction for binaural hearing aids, (Vancouver, 2013), pp. 8648–8652. doi:10.1109/ICASSP.2013.6639354
19. A Kuklasinski, S Doclo, SH Jensen, J Jensen, in *Proceedings of the European Signal Processing Conference (EUSIPCO)*. Maximum likelihood based multi-channel isotropic reverberation reduction for hearing aids, (Lisbon, 2014)
20. J Bitzer, KU Simmer, in *Microphone Arrays*, ed. by M Brandstein, D Ward. Chapter 2: superdirective microphone arrays (Springer, Berlin, 2010), pp. 19–38
21. BD Van Veen, KM Buckley, Beamforming: a versatile approach to spatial filtering. IEEE ASSP Mag. **5**(2), 4–24 (1988). doi:10.1109/53.665
22. S Gannot, D Burshtein, E Weinstein, Signal enhancement using beamforming and nonstationarity with applications to speech. IEEE Trans. Signal Process. **49**(8), 1614–1626 (2001). doi:10.1109/78.934132
23. S Braun, EAP Habets, in *Proceedings of the European Signal Processing Conference (EUSIPCO)*. Dereverberation in noisy environments using reference signals and a maximum likelihood estimator, (2013), pp. 1–5
24. KU Simmer, J Bitzer, C Marro, in *Microphone Arrays*, ed. by M Brandstein, D Ward. Chapter 3: post-filtering techniques (Springer, Berlin, 2001), pp. 39–60
25. U Kjems, J Jensen, in *Proceedings of the European Signal Processing Conference (EUSIPCO)*. Maximum likelihood based noise covariance matrix estimation for multi-microphone speech enhancement, (Bucharest, 2012), pp. 295–299
26. H Ye, RD DeGroat, Maximum likelihood DOA estimation and asymptotic Cramér-Rao bounds for additive unknown colored noise. IEEE Trans. Signal Process. **43**(4), 938–949 (1995). doi:10.1109/78.376846
27. M Cooke, A glimpsing model of speech perception in noise. J. Acoust. Soc. Am. **119**(3), 1562–1573 (2006). doi:10.1121/1.2166600
28. DS Brungart, PS Chang, BD Simpson, D Wang, Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. J. Acoust. Soc. Am. **120**(6), 4007–4018 (2006). doi:10.1121/1.2363929
29. J Thiemann, M Müller, S van de Par, in *Proceedings of the European Signal Processing Conference (EUSIPCO)*. A binaural hearing aid speech enhancement method maintaining spatial awareness for the user, (Lisbon, 2014), pp. 321–325
30. TJ Klasen, T Van den Bogaert, M Moonen, J Wouters, Binaural noise reduction for hearing aids that preserve interaural time delay cues. IEEE Trans. Signal Process. **55**(4), 1579–1585 (2007). doi:10.1109/TSP.2006.888897

31. D Marquardt, Development and evaluation of psychoacoustically motivated binaural noise reduction and cue preservation techniques. PhD thesis, Fakultät für Medizin und Gesundheitswissenschaften, Carl von Ossietzky Universität Oldenburg. (2015)
32. H Kayser, SD Ewert, J Anemüller, T Rohdenburg, V Hohmann, B Kollmeier, Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses. EURASIP J. on Applied Sig. Proc (2009). doi:10.1155/2009/298605
33. K Wagener, V Kühnel, B Kollmeier, Entwicklung und Evaluation eines Satztests für die deutsche Sprache, I: Design des Oldenburger Satztests. Zeitschrift für Audiologie. **38**(1), 4–15 (1999)
34. K Wagener, T Brand, B Kollmeier, Entwicklung und Evaluation eines Satztests für die deutsche Sprache, II: Optimierung des Oldenburger Satztests. Zeitschrift für Audiologie. **38**(2), 44–56 (1999)
35. K Wagener, T Brand, B Kollmeier, Entwicklung und Evaluation eines Satztests für die deutsche Sprache, III: Evaluation des Oldenburger Satztests. Zeitschrift für Audiologie. **38**(3), 86–95 (1999)
36. T Van den Bogaert, S Doclo, J Wouters, M Moonen, Speech enhancement with multichannel Wiener filter techniques in multimicrophone binaural hearing aids. J. Acoust. Soc. Am. **125**(1), 360–371 (2009). doi:10.1121/1.3023069
37. JE Greenberg, PM Peterson, PM Zurek, Intelligibility-weighted measures of speech-to-interference ratio and speech system performance. J. Acoust. Soc. Am. **94**(5), 3009–3010 (1993). doi:10.1121/1.407334
38. International Telecommunication Union, *ITU-T Recommendation P.862, Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs*, (Geneva, 2001)
39. M Dietz, SD Ewert, V Hohmann, Auditory model based direction estimation of concurrent speakers from binaural signals. Speech Comm. **53**(5), 592–605 (2011). doi:10.1016/j.specom.2010.05.006
40. D Marquardt, V Hohmann, S Doclo, Interaural coherence preservation in multi-channel wiener filtering-based noise reduction for binaural hearing aids. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(12), 2162–2176 (2015). doi:10.1109/TASLP.2015.2471096
41. International Telecommunication Union, *ITU-R Recommendation BS.1534-1, Method for the subjective assessment of intermediate quality level of coding systems*, (Geneva, 2003)
42. *International Organization for Standardization*, ISO Standard 8253-3:2012 Acoustics-audiometric test methods-part 3: speech audiometry (2012)
43. U Kjems, MS Pedersen, JB Boldt, T Lunner, D Wang, in *Proceedings of the European Signal Processing Conference (EUSIPCO)*. Speech intelligibility of ideal binary masked mixtures, (Aalborg, 2010), pp. 1909–1913