Azarpour and Enzner *EURASIP Journal on Advances in Signal Processing* (2017) 2017:49
DOI 10.1186/s13634-017-0485-9

EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

Binaural noise reduction via cue-preserving MMSE filter and adaptive-blocking-based noise PSD estimation

Masoumeh Azarpour* and Gerald Enzner

Abstract

Binaural noise reduction, with applications for instance in hearing aids, has been a very significant challenge. This task relates to the optimal utilization of the available microphone signals for the estimation of the ambient noise characteristics and for the optimal filtering algorithm to separate the desired speech from the noise. The additional requirements of low computational complexity and low latency further complicate the design. A particular challenge results from the desired reconstruction of binaural speech input with spatial cue preservation. The latter essentially diminishes the utility of multiple-input/single-output filter-and-sum techniques such as beamforming. In this paper, we propose a comprehensive and effective signal processing configuration with which most of the aforementioned criteria can be met suitably. This relates especially to the requirement of efficient online adaptive processing for noise estimation and optimal filtering while preserving the binaural cues. Regarding noise estimation, we consider three different architectures: interaural (ITF), cross-relation (CR), and principal-component (PCA) target blocking. An objective comparison with two other noise PSD estimation algorithms demonstrates the superiority of the blocking-based noise estimators, especially the CR-based and ITF-based blocking architectures. Moreover, we present a new noise reduction filter based on minimum mean-square error (MMSE), which belongs to the class of common gain filters, hence being rigorous in terms of spatial cue preservation but also efficient and competitive for the acoustic noise reduction task. A formal real-time subjective listening test procedure is also developed in this paper. The proposed listening test enables a real-time assessment of the proposed computationally efficient noise reduction algorithms in a realistic acoustic environment, e.g., considering time-varying room impulse responses and the Lombard effect. The listening test outcome reveals that the signals processed by the blocking-based algorithms are significantly preferred over the noisy signal in terms of instantaneous noise attenuation. Furthermore, the listening test data analysis confirms the conclusions drawn based on the objective evaluation.

Keywords: Equalization-cancelation, Noise estimation, Cue preservation, Binaural noise reduction, Real-time listening test

1 Introduction

Hearing loss is a common sensory deficiency, as reported, e.g., in [1]. Thus, hearing technologies should provide a remarkable compensation of hearing deficits for people with hearing loss. For instance, modern hearing aids utilize a variety of techniques to enhance the quality and intelligibility of the desired signal in the presence of ambient noise. However, noise reduction generally is seen as a difficult task, and the respective performance still remains quite limited in realistic scenarios.

Noise reduction algorithms can be categorized in different ways. The number of employed microphones is a criterion used to classify such algorithms into single-channel, dual-channel/binaural, and multi-channel algorithms. In this study, we will address the binaural noise reduction problem where the left and right microphone signals interact to deliver a reliable noise reduction performance. In contrast, bilateral signal processing refers to the treatment of the left and right ear independently. Here, the binaural cues, which are particularly important for sound

*Correspondence: masoumeh.azarpour@rub.de
Ruhr-Universität Bochum, Institute of Communication Acoustics, Universitätsstr. 150, Bochum, Germany

localization, will be distorted. It has been reported in [2] that if the noise reduction methods embedded in hearing aids do not preserve the binaural cues, hearing-impaired people prefer to disable the noise reduction option in their hearing aids for the sake of better sound localization.

The preservation of the binaural cues, particularly the interaural level difference (ILD) and the interaural time difference (ITD), is an important issue that needs to be treated properly in binaural signal processing in addition to noise reduction and speech preservation. Thus, different noise reduction techniques have been proposed to suppress noise while the spatial impression of the desired and interference sources are kept undistorted. These techniques can be effectively dichotomized into two main categories.

The first category mostly consists of multichannel algorithms, therein combining spatial and spectral filtering, which attempt to reduce noise with an additional constraint on auditory scene preservation [3–5]. These algorithms are commonly designed by modifying the noise-reduction-related cost functions such that the binaural cues are kept undistorted [3, 6, 7]. It has been shown that the binaural multichannel Wiener filter (MWF) [8] and the binaural minimum-variance distortionless-response (MVDR) beamformer [9, 10] can preserve the binaural cues of the speech components, whereas the binaural cues of the noise components will be distorted. To preserve the binaural cues of a directional noise source, the authors in [11] introduced a new parameter in MWF to facilitate a trade off between noise reduction and noise binaural cue preservation. Another extension of MWF with partial noise estimation was proposed in [12, 13]. In [14], a term related to the interaural transfer function of the noise source was integrated into the noise reduction cost function to preserve the binaural cues of the noise source (MWF-ITF). Later, a simplified MWF-ITF was proposed in [7] and offers a closed-form solution for binaural noise reduction and noise cue preservation. Moreover, additional linear constraints have been considered in the MVDR beamformer [10, 15] and the binaural MWF [16, 17] with the aim of preserving the binaural cues of an interfering source. Nevertheless, the techniques discussed so far are not well suited for the spatial preservation of diffuse noise. To preserve the interaural coherence (IC) of the residual noise components of diffuse noise, the binaural MWF is extended using additional IC-related cost functions [18–21].

The second category of noise reduction techniques includes algorithms that employ a real-valued common spectral gain function [22–24]. The interfering signal, including the ambient noise and reverberation, is assumed to be spatially diffuse. Applying the zero-phase common function to the signals of the left and right ears ensures the preservation of the binaural cues. The common spectral gain function can be obtained by either minimizing the spectral distance between the bilateral gain functions [25] or computing the compound of the bilateral gains heuristically [26–29]. For instance, [26] exploits the minimum, maximum, and average of two independent single-channel gain functions at the left and right ears to derive a common gain. In this work, the minimum of the bilateral gains in each frame and frequency bin was considered to be the most efficient. The aforementioned common spectral gain functions are conventionally adopted from single-channel techniques. Therefore, they often suffer from low noise reduction and potential speech artifacts, although they can provide the perfect preservation of spatial impressions. The suggested solutions are mostly developed by heuristically combing the single-channel gain functions and hence are not necessarily optimal. The concept of a common spectral noise reduction filter is also frequently found in the form of a spectral postfilter to MVDR beamformers. In the postfiltering scheme, the Wiener filter based on the mean-square-error (MSE) criteria [30, 31] is often the starting point for variations and modifications, e.g., [32–34]. For instance, in [35], a common spectral gain function controlled by a superdirective beamformer design based on a head-related transfer function (HRTF) model was developed.

Different assumptions on noise statistics lead to various optimal filter coefficients. For instance, Zelinski's spectral postfilter [36] is derived assuming uncorrelated noise in the channels. This assumption, however, has been generalized to a low-frequency coherent noise using the coherence model of spherically ideal diffuse noise [37]. Later, the authors in [28] proposed to take the average of the left and right bilateral filters as a post-filter for dual-channel noise reduction, where the ambient noise signals are assumed to be spatially uncorrelated. It can be shown that this averaging leads to a realization of Zelinski's filter provided that the noises received at the microphones are uncorrelated and have identical power at all frequencies.

In many speech enhancement algorithms, such as Wiener filtering, prior knowledge of the noise statistics is a prerequisite for successful ambient noise reduction [30, 38]. Recently, the target cancelation technique has been employed in noise power estimation. For instance, it has been proposed to use the blind source separation (BSS) approach for canceling the target speech components in a diffuse noise field and consequently to estimate the noise power at the output of the blocking system [39]. Later, the same approach was employed in [40] to estimate the reverberation tail, which is considered as diffuse noise. A spectral correction gain function based on the BSS de-mixing matrix was derived to reduce the bias of the estimated noise PSD. In [41], we proposed a binaural noise PSD estimator based on the equalization-cancelation technique. The target speech

signal is equalized and canceled by two independent least-mean-square (LMS)-type algorithms for the left and right noise PSD estimation. A correction gain is then derived using the estimated interaural transfer functions between the left and right ears. In [42], we proposed to employ a blind system identification approach based on the cross-relation error minimization to estimate the noise PSD using the cross-relation residual. The successful application of the estimated noise power for speech enhancement was initially demonstrated in [41, 42] with hearing aid application.

In this contribution, a new binaural cue-preserving noise reduction filter, yet based on the MMSE criteria, is proposed (Fig. 1). The proposed noise reduction filter possesses properties such as optimality and ease of implementation. Based on a common gain function, the mean-square error is rigorously minimized jointly in the left and right ear, thereby delivering optimal noise reduction with exact binaural cue preservation of the target speech and residual noise.

To implement the proposed cue-preserving MMSE filter, this paper further investigates and compares a broad range of subspace techniques for noise PSD estimation. This includes the interaural transfer function blocking-based noise PSD estimator (ITFB) [41] (Fig. 2a) and the cross-relation-based noise PSD estimators (CRB) [42] (Fig. 2b), which were previously evaluated under anechoic conditions. They are evaluated here in a more realistic acoustic environment. The comparison is conducted in an ambient noise environment with moderate reverberation such as in a cafeteria, outdoor street, or congress environment. Additionally, a new noise power estimation based on speech blocking is investigated (PCAB, Fig. 2c). That algorithm employs adaptive principle component analysis (PCA) [43]. The adaptive PCA was previously used for the blind channel identification and equalization in hearing aids [44]. The speech components are canceled in the error signals of the adaptive PCA-based blocking. A spectral correction gain derived using the estimated impulse responses and the noise coherence is then applied to correct the biased noise components remaining in the blocking output.

In this paper, additionally, we develop a real-time subjective listening test for the evaluation of binaural noise reduction algorithms. The developed listening test exhibits remarkable benefits for a valid assessment of noise reduction algorithms such as (1) realistic exposure to speech and noise; (2) natural speech performance, e.g., including the Lombard effect [45]; (3) different signal-to-noise ratios (SNRs) and noise types (sensor noise, ambient noise, and reverberation); and (4) easy variations in spatial cues.

The remainder of this paper is organized as follows. In Section 2, we formulate the binaural signal model and the noise reduction problem. The proposed binaural cue-preserving MMSE filter is introduced in Section 3. Section 4 presents the theory of subspace noise estimation, and Section 5 introduces the instrumental evaluation tools related to adaptive target blocking. In Section 6, the performance of the proposed algorithms is evaluated in terms of impulse response estimation, noise PSD estimation, noise tracking, and speech enhancement. Finally, Section 7 is devoted to the developed real-time listening test and subjective evaluation of proposed blocking-based algorithms.

2 Binaural signal model

Let $y_i(k)$, with $i \in \{r, l\}$, denote the binaural microphone signals at sampling time index k, which can be expressed as

$$y_i(k) = \sum_{n=0}^{\infty} s(k-n)h_i(n) + n_{i,\mathrm{a}}(k), \tag{1}$$

where $s(k)$, $h_i(k)$, and $n_{i,\mathrm{a}}(k)$ are the target speech, the binaural room impulse responses (BRIR), and the ambient

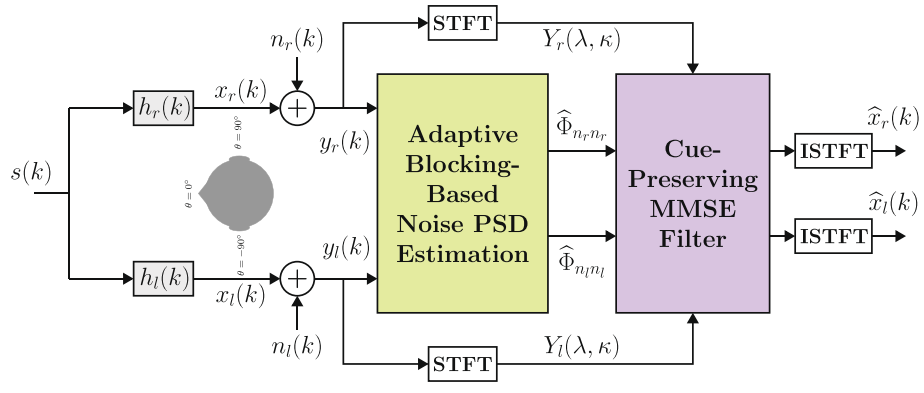

Fig. 1 Schematic block diagram of the proposed binaural noise reduction system

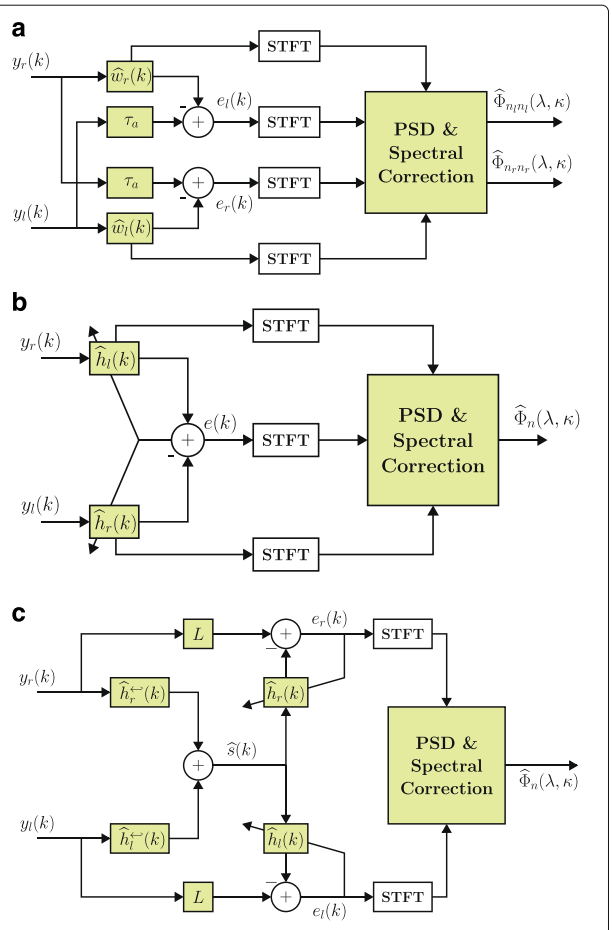

Fig. 2 Schematic block diagram of **a** ITF-blocking-based noise PSD estimation, **b** CR-blocking-based noise PSD estimation, and **c** PCA-blocking-based noise PSD estimation

background noises, respectively. In this study, we used moderately reverberant BRIRs. Thus, the clean speech signal can be decomposed into the desired direct sound and early reflection part, $n = 0 \ldots L$, and the undesired reverberation components $n = L+1 \ldots \infty$,

$$\begin{aligned} y_i(k) &= \sum_{n=0}^{L} s(k-n)h_i(n) \\ &\quad + \sum_{n=L+1}^{\infty} s(k-n)h_i(n) + n_{i,\mathrm{a}}(k), \\ &= x_i(k) + n_i(k), \end{aligned} \tag{2}$$

where the effective noise $n_i(k)$ consists of the moderate reverberation and the ambient noise $n_{i,\mathrm{a}}(k)$. The vectors $\mathbf{y}_i(k) = \left[y_i(k)\, y_i(k-1) \ldots y_i(k-L+1)^T\right]$ of L successive samples are also used, where the superscript $(.)^T$ denotes the vector transposition. The other signal vectors, e.g., $\mathbf{x}_i(k)$ and $\mathbf{n}_i(k)$, are defined in the same way as $\mathbf{y}_i(k)$; thus, $\mathbf{y}_i(k) = \mathbf{x}_i(k) + \mathbf{n}_i(k)$. The short-time Fourier transform (STFT) [46] of (2) reads

$$Y_i(\lambda,\kappa) = X_i(\lambda,\kappa) + N_i(\lambda,\kappa) \tag{3}$$

where $\lambda = 0, \ldots, M$ and $\kappa \in \mathbf{Z}$ indicate the frequency bin and frame indices, respectively.

The desired speech components $\widehat{X}_i$ are then retrieved in the MSE sense by applying an optimal filter $G(\lambda,\kappa)$ to the noisy signal,

$$\widehat{X}_i(\lambda,\kappa) = G(\lambda,\kappa)Y_i(\lambda,\kappa), \tag{4}$$

which will be elaborated upon further in the next section. The time-varying power spectral densities (PSDs) of the noise and the noisy signals are defined as $\Phi_{n_i n_i}(\lambda,\kappa) = E\left\{|N_i(\lambda,\kappa)|^2\right\}$, and $\Phi_{y_i y_i}(\lambda,\kappa) = \mathrm{E}\left\{|Y_i(\lambda,\kappa)|^2\right\}$, respectively, where E{.} denotes the statistical expectation operator. We use a first-order recursive system,

$$\widehat{\Phi}_{y_i y_i}(\lambda,\kappa) = \alpha\widehat{\Phi}_{y_i y_i}(\lambda,\kappa-1) + (1-\alpha)\,|Y_i(\lambda,\kappa)|^2, \tag{5}$$

to estimate the auto-PSDs of the accessible signals with a smoothing factor of $0 \leq \alpha < 1$. The cross PSDs are estimated analogously. For the sake of simplicity, the frequency index λ and frame index κ will be omitted hereafter unless they are needed for clarity. The enhanced signals $\widehat{X}_i(\lambda,\kappa)$ are then transferred back to the time domain by applying the inverse STFT and employing the overlap-add (OLA) technique [47].

3 Binaural cue-preserving MMSE filter

In the following, we present a binaural cue-preserving filter based on the MMSE criterion. The noise reduction problem is to find a statistically optimal filter G_o that jointly minimizes

$$\begin{aligned} \mathcal{J}(G) = \mathrm{E}\big\{&|X_l(\lambda,\kappa) - G(\lambda,\kappa)Y_l(\lambda,\kappa)|^2 \\ &+ |X_r(\lambda,\kappa) - G(\lambda,\kappa)Y_r(\lambda,\kappa)|^2\big\} \end{aligned} \tag{6}$$

such that the optimal filter is

$$G_o = \underset{G}{\mathrm{argmin}}(\mathcal{J}(G)). \tag{7}$$

Assuming that the noise and speech signals are uncorrelated, i.e., $\Phi_{x_l y_l} = \Phi_{x_l x_l}$, the cost function simplifies as

$$\begin{aligned} \mathcal{J}(G) = \Phi_{x_l x_l} &+ |G|^2\Phi_{y_l y_l} - 2\Phi_{x_l x_l}G \\ &+ \Phi_{x_r x_r} + |G|^2\Phi_{y_r y_r} - 2\Phi_{x_r x_r}G. \end{aligned} \tag{8}$$

By taking the derivative of (8) with respect to real G,

$$\frac{\partial \mathcal{J}(G)}{\partial G} = G\Phi_{y_l y_l} - \Phi_{x_l x_l} + G\Phi_{y_r y_r} - \Phi_{x_r x_r}, \tag{9}$$

and equating the result to zero, the frequency response of our proposed binaural cue-preserving MMSE filter reads [48]

$$G_o = \frac{\Phi_{x_l x_l} + \Phi_{x_r x_r}}{\Phi_{y_l y_l} + \Phi_{y_r y_r}} = 1 - \frac{\Phi_{n_l n_l} + \Phi_{n_r n_r}}{\Phi_{y_l y_l} + \Phi_{y_r y_r}}. \tag{10}$$

To attenuate musical noise introduced in the enhanced signal and to balance the noise reduction and speech distortion, an over-subtraction factor $\beta \geq 1$ [30] is employed, and the filters are spectrally floored to $G_{\min}$, i.e.,

$$G_o = \max\left(1 - \beta\frac{\Phi_{n_l n_l} + \Phi_{n_r n_r}}{\Phi_{y_l y_l} + \Phi_{y_r y_r}}, G_{\min}\right). \tag{11}$$

4 Noise PSD estimation via adaptive speech blocking

The improvement in the speech quality and intelligibility depends remarkably on the accuracy of the noise power estimate. The estimators presented here are inspired by the target cancelation technique, in which the coherent target speech signal is blocked from the microphone signals to retrieve the noise components. However, the estimated noise components at the output of the blocking system are always the filtered versions of the actual noise signal. A spectral correction gain, obtained via the estimated blocking filters, is thus employed in each case to undo this filtering effect.

It should also be mentioned that the assumption of target speech cancelation would not be completely fulfilled in the presence of the observation noise, which is the case considered in this paper. Therefore, the residual speech components (called speech leakage) leak into the estimated noise, increasing the estimated noise power and possibly leading to speech distortion in the enhancement stage of Fig. 1. The speech leakage problem in blocking-based-noise PSD estimators will be elaborated upon more precisely in Section 6.2 of this paper.

The algorithms that will be elaborated upon in this section are all based on square-error minimization. However, the filter structures are different for each method, c.f., Fig. 2a, b, and c. All methods can be understood as being different forms of subspace analysis, with different origins in the signal or noise-subspace analysis; however, they will all be cast into the common framework of a noise PSD estimator here.

4.1 ITF-based adaptive blocking (ITFB)

The interaural transfer function (ITF) estimation errors, subject to minimization, are written as [41]

$$\begin{aligned} e_l(k) &= y_l(k-\tau_a) - \widehat{\mathbf{w}}_r^T(k)\mathbf{y}_r(k), \\ e_r(k) &= y_r(k-\tau_a) - \widehat{\mathbf{w}}_l^T(k)\mathbf{y}_l(k), \end{aligned} \tag{12}$$

where the causality delay of τ_a has been added to ensure that the system identification problem is causal. The left-to-right and right-to-left interaural impulse responses $\widehat{\mathbf{w}}_i$, with $i \in \{l, r\}$, are then updated iteratively according to

$$\begin{aligned} \widehat{\mathbf{w}}_l(k+1) &= \widehat{\mathbf{w}}_l(k) + \mu_l(k)e_r(k)\mathbf{y}_l(k), \\ \widehat{\mathbf{w}}_r(k+1) &= \widehat{\mathbf{w}}_r(k) + \mu_r(k)e_l(k)\mathbf{y}_r(k), \end{aligned} \tag{13}$$

where $\mu_i(k) = \mu_0/\mathbf{y}_i^T(k)\mathbf{y}_i(k)$ is the normalized stepsize with a fixed stepsize of $0 < \mu_0 \leq 1$. This minimization of the respective error signal powers is in accordance with the sample-based normalized least-mean-square (NLMS) algorithm as shown here in the time domain or alternatively via the more efficient frequency-domain adaptive filter (FDAF) [49]. In either case, two parallel adaptive filters are implemented to perform the minimization of the left and right error signals independently. The presence of observation noise will naturally affect the adaptive filter performance, but we will rely on the general insight that the target cancelation error of LMS-type adaptive filters is theoretically several dB below the observation noise level [30, 44]. Although the actual target cancelation error depends on the stepsize of the LMS algorithm, we found that the range of stepsize factors $0.01 < \mu < 0.1$ to be sufficient to deduce an accurate noise PSD estimation from the error signal of the adaptive filters. With this argument, we can characterize the error signals of (12) as

$$\begin{aligned} e_i(k) &= x_i(k-\tau_a) + n_i(k-\tau_a) \\ &\quad - \widehat{\mathbf{w}}_j^T(k)\mathbf{x}_j(k) - \widehat{\mathbf{w}}_j^T(k)\mathbf{n}_j(k), \\ &\approx n_i(k-\tau_a) - \widehat{\mathbf{w}}_j^T(k)\mathbf{n}_j(k), \quad i \neq j \in \{l, r\}. \end{aligned} \tag{14}$$

By computing the PSDs of the error signals according to (5), a system of equations including the left and right noise PSDs is obtained,

$$\begin{aligned} \widehat{\Phi}_{e_l e_l} &= \Phi_{n_l n_l} + \left|\widehat{W}_r\right|^2 \Phi_{n_r n_r} - 2\mathrm{Re}\left\{e^{j\frac{2\pi}{M}\lambda\tau_a}\widehat{W}_r\Phi_{n_l n_r}\right\}, \\ \widehat{\Phi}_{e_r e_r} &= \Phi_{n_r n_r} + \left|\widehat{W}_l\right|^2 \Phi_{n_l n_l} - 2\mathrm{Re}\left\{e^{j\frac{2\pi}{M}\lambda\tau_a}\widehat{W}_l\Phi_{n_l n_r}\right\}, \end{aligned} \tag{15}$$

with an STFT length of M. The PSD of the left and right noise signals, $\widehat{\Phi}_{n_l n_l}$ and $\widehat{\Phi}_{n_r n_r}$, respectively, can then be derived by solving the simultaneous equations in (15), and consequently, the noise distortion due to the blocking filters can be corrected. In this process, at least three different noise coherence models can be assumed: (1) uncorrelated noise, (2a) free-field spherically isotropic diffuse noise, and (2b) measured or semi-analytical head-related coherence.

4.1.1 Uncorrelated noise

First, we assume that the noise signals in the left and right microphone are uncorrelated $\Phi_{n_l n_r} = \Phi_{n_r n_l} = 0$ which is a reasonable assumption for a diffuse noise field above a cutoff frequency. Therefore, (15) will be a system of linear equations. By solving the equations, the PSDs of the left and right noise signals can be derived as

$$\widehat{\Phi}_{n_l n_l} = \frac{\widehat{\Phi}_{e_l e_l} - \left|\widehat{W}_r\right|^2 \widehat{\Phi}_{e_r e_r}}{1 - \left|\widehat{W}_l\right|^2 \left|\widehat{W}_r\right|^2}, \qquad (16)$$

$$\widehat{\Phi}_{n_r n_r} = \frac{\widehat{\Phi}_{e_r e_r} - \left|\widehat{W}_l\right|^2 \widehat{\Phi}_{e_l e_l}}{1 - \left|\widehat{W}_l\right|^2 \left|\widehat{W}_r\right|^2}.$$

Many practical noise signals exhibit high correlation in the low-frequency range. Therefore, the premise that the noise signal in real acoustic scenarios is fully uncorrelated is not true. Thus, the proposed solution with the assumption of an uncorrelated noise model indeed leads to noise PSD underestimation at low frequencies where the noise signals are correlated (not shown here). The low-frequency compensation of the noise PSD will be addressed in the following section.

4.1.2 Diffuse noise

To overcome the underestimation of the noise power at low frequencies, we employ the noise coherence function. The complex coherence between two noise signals is generally defined as [50]

$$\Gamma_{n_l n_r}(\lambda, \kappa) = \frac{\Phi_{n_l n_r}(\lambda, \kappa)}{\sqrt{\Phi_{n_l n_l}(\lambda, \kappa)\Phi_{n_r n_r}(\lambda, \kappa)}}, \qquad (17)$$

where $\Phi_{n_i n_j}(\lambda, \kappa), i, j \in \{l, r\}$ are the cross and auto-PSD of the noise signals, which can be estimated using a first-order recursive equation as in (5) when $n_l(k)$ and $n_r(k)$ are available. Substituting (17) into (15) will lead to a nonlinear system of equations. To simplify the equations, the noise PSDs at the left and right ear are considered to be equal. In [41], it was shown that for measured noise signals, the assumptions of equal noise PSDs at the two microphones are more plausible at low frequencies than at high frequencies. Assuming equal noise PSDs, i.e., $\Phi_{n_l n_l} = \Phi_{n_r n_r} = \Phi_n$ at the two microphones, the cross PSD, $\Phi_{n_l n_r}$ in (15), consequently can be expressed based on the left and right noise PSDs and the coherence function, i.e., $\Phi_{n_r n_l} = \Phi_{n_l n_r} = \Gamma_{n_l n_r}\Phi_n$, therein considering that the noise coherence of a diffuse noise field is real valued. Therefore, the noise PSD estimates can be obtained as

$$\widehat{\Phi}_{n_l n_l} = \frac{\widehat{\Phi}_{e_l}}{1 + \left|\widehat{W}_r\right|^2 - 2\mathrm{Re}\left\{e^{j\frac{2\pi}{M}\lambda\tau_a}\widehat{W}_r\Gamma_{n_l n_r}\right\}}, \qquad (18)$$

$$\widehat{\Phi}_{n_r n_r} = \frac{\widehat{\Phi}_{e_r}}{1 + \left|\widehat{W}_l\right|^2 - 2\mathrm{Re}\left\{e^{j\frac{2\pi}{M}\lambda\tau_a}\widehat{W}_l\Gamma_{n_l n_r}\right\}}.$$

A spectral flooring of −20 dB is additionally used in the denominator to avoid division by zero. Moreover, the following noise coherence models can be considered here: (1) free-field diffuse noise coherence, (2) the head-related coherence model [51], and (3) head-related coherence estimates. It has been observed that an accurate estimation of the noise PSD can be obtained if a good model of the noise coherence is employed. Therefore, we suggest using the 2D head-related coherence model proposed in [51].

4.2 CR-based adaptive blocking (CRB)

The cross-relation (CR) error between the microphone signal is given as, for instance [44],

$$e(k) = \widehat{\mathbf{h}}_r^T(k)\mathbf{y}_l(k) - \widehat{\mathbf{h}}_l^T(k)\mathbf{y}_r(k), \qquad (19)$$

where the left and right impulse responses $\widehat{\mathbf{h}}_i(k) = \left[\widehat{h}_i(0)\ \widehat{h}_i(1)\ \ldots\ \widehat{h}_i(L-1)\right]^T$ can be determined by a stereo normalized least-mean-square (NLMS) algorithm [42, 44]:

$$\widehat{\mathbf{h}}_l(k+1) = \widehat{\mathbf{h}}_l(k) + \mu(k)e(k)\mathbf{y}_r(k), \qquad (20)$$

$$\widehat{\mathbf{h}}_r(k+1) = \widehat{\mathbf{h}}_r(k) - \mu(k)e(k)\mathbf{y}_l(k),$$

where the normalized stepsize

$$\mu(k) = \mu_0\left(\mathbf{y}_l^T(k)\mathbf{y}_l(k) + \mathbf{y}_r^T(k)\mathbf{y}_r(k)\right)^{-1} \qquad (21)$$

governs the convergence rate of the algorithm.

The estimated impulse responses are further normalized to unit norm in each iteration of the recursive adaptation, i.e.,

$$\widehat{\mathbf{h}}_l^T(k)\widehat{\mathbf{h}}_l(k) + \widehat{\mathbf{h}}_r^T(k)\widehat{\mathbf{h}}_r(k) = 1, \qquad (22)$$

to avoid trivial solutions. Substituting the binaural signal model (1) into (19), we have

$$e(k) = \widehat{\mathbf{h}}_r^T(k)(\mathbf{x}_l(k) + \mathbf{n}_l(k)), \qquad (23)$$
$$- \widehat{\mathbf{h}}_l^T(k)(\mathbf{x}_r(k) + \mathbf{n}_r(k)).$$

Because we expect that $\widehat{\mathbf{h}}_r^T(k)\mathbf{x}_l(k) \approx \widehat{\mathbf{h}_l^T}(k)\mathbf{x}_r(k)$ after the error signal minimization in cross-relation techniques, the speech related part in (23) is canceled. Even when the estimated channels are altered by an unknown yet common convolutive operation, i.e., $\widehat{h}_i(k) = f(k) * h_i(k)$ [52], the common convolutive error, which might be a drawback in blind channel identification, does not seriously affect the speech blocking performance because it applies simultaneously to both the left and right estimated impulse responses. Therefore, the error signal

$$e(k) \approx \widehat{\mathbf{h}}_r^T(k)\mathbf{n}_l(k) - \widehat{\mathbf{h}}_l^T(k)\mathbf{n}_r(k), \qquad (24)$$

contains the filtered noise components of the left and right microphone signals. Thus, although the error signal can be considered as an estimation of the noise signal, this estimation is biased because the left and right noise

signal components are filtered by the estimated impulse responses. Transferring (24) into the PSD domain, we obtain

$$\widehat{\Phi}_e = \left|\widehat{H}_r\right|^2 \Phi_{n_l n_l} + \left|\widehat{H}_l\right|^2 \Phi_{n_r n_r} - 2\mathrm{Re}\left\{\widehat{H}_l \widehat{H}_r^* \Phi_{n_l n_r}\right\}. \tag{25}$$

Moreover, the left and right noise PSDs are again assumed to be identical to solve the single Eq. (25), i.e., $\Phi_{n_r n_r} = \Phi_{n_l n_l} = \Phi_n$. The cross PSD of the left and right noise signals is again replaced by the coherence of the noise signals, i.e., $\Phi_{n_l n_r} = \Phi_n \Gamma_{n_l n_r}$. Thus,

$$\widehat{\Phi}_e = \left|\widehat{H}_r\right|^2 \Phi_n + \left|\widehat{H}_l\right|^2 \Phi_n - 2\mathrm{Re}\left\{\widehat{H}_l \widehat{H}_r^* \Gamma_{n_l n_r} \Phi_n\right\}. \tag{26}$$

The error PSD $\widehat{\Phi}_e$ is obtained using the first-order recursive averaging according to (5), with $E(\lambda, \kappa)$ being the STFT of the cross-relation error signal $e(k)$ according to (19). By solving (26), the estimated noise PSD is obtained as

$$\widehat{\Phi}_n = \frac{\widehat{\Phi}_e}{\left|\widehat{H}_r\right|^2 + \left|\widehat{H}_l\right|^2 - 2\mathrm{Re}\left\{\widehat{H}_l \widehat{H}_r^* \Gamma_{n_l n_r}\right\}}. \tag{27}$$

To avoid division by zero, a spectral flooring is applied to limit the denominator to −20 dB.

4.3 PCA-based adaptive blocking (PCAB)

In this algorithm, the left and the right source-to-microphone transfer functions are identified by minimizing the error signal between microphone signal and an estimated source signal, i.e., $\widehat{s}(k)$ [43, 44, 53]

$$\begin{aligned} e_l(k) &= y_l(k-L) - \widehat{\mathbf{h}}_l^T \widehat{\mathbf{s}}(k), \\ e_r(k) &= y_r(k-L) - \widehat{\mathbf{h}}_r^T \widehat{\mathbf{s}}(k). \end{aligned} \tag{28}$$

The estimated source signal $\widehat{\mathbf{s}}(k)$ is a vector of L recent successive samples $\widehat{\mathbf{s}}(k) = [\widehat{s}(k)\, \widehat{s}(k-1) \ldots \widehat{s}(k-L+1)]^T$ resulting in a matched filter operation,

$$\widehat{s}(k) = \widehat{\mathbf{h}}_l^{T\hookleftarrow}(k) \mathbf{y}_l(k) + \widehat{\mathbf{h}}_r^{T\hookleftarrow}(k) \mathbf{y}_r(k), \tag{29}$$

where $(.)^{\hookleftarrow}$ denotes the time-reversed estimated impulse response. The estimated left and right impulse responses are updated according to the LMS style,

$$\begin{aligned} \widehat{\mathbf{h}}_l(k+1) &= \widehat{\mathbf{h}}_l(k) + \mu(k) e_l(k) \widehat{\mathbf{s}}(k), \\ \widehat{\mathbf{h}}_r(k+1) &= \widehat{\mathbf{h}}_r(k) + \mu(k) e_r(k) \widehat{\mathbf{s}}(k). \end{aligned} \tag{30}$$

We can transfer (28) into the STFT domain,

$$\begin{aligned} E_l(\kappa, \lambda) &= e^{-j\frac{2\pi}{M}\lambda L} Y_l(\kappa, \lambda) - \widehat{H}_l(\kappa, \lambda) \widehat{S}(\kappa, \lambda), \\ E_r(\kappa, \lambda) &= e^{-j\frac{2\pi}{M}\lambda L} Y_r(\kappa, \lambda) - \widehat{H}_r(\kappa, \lambda) \widehat{S}(\kappa, \lambda), \end{aligned} \tag{31}$$

and the matched filter output of (29) is

$$\widehat{S} = e^{-j\frac{2\pi}{M}\lambda L} \widehat{H}_l^* Y_l + e^{-j\frac{2\pi}{M}\lambda L} \widehat{H}_r^* Y_r. \tag{32}$$

Assuming the proper transfer function estimation, i.e., $\widehat{H}_i = H_i F$, where F is a common filter error [52], (32) is expressed as

$$\begin{aligned} \widehat{S} = {} & e^{-j\frac{2\pi}{M}\lambda L} S F^{-1} \left(\left|\widehat{H}_l\right|^2 + \left|\widehat{H}_r\right|^2\right) \\ & + e^{-j\frac{2\pi}{M}\lambda L} N_l \widehat{H}_l^* + e^{-j\frac{2\pi}{M}\lambda L} N_r \widehat{H}_r^*. \end{aligned} \tag{33}$$

Because the recursive algorithm in (30) can be observed as a one-to-one translation of a frequency-domain (bin-wise) representation of adaptive PCA [54], it provides approximately a bin-wise unit norm, i.e., $\left|\widehat{H}_l\right|^2 + \left|\widehat{H}_r\right|^2 \approx 1$ when the convergence toward the principle components is achieved. Thus,

$$\widehat{S} = e^{-j\frac{2\pi}{M}\lambda L} \left(F^{-1} S + N_l \widehat{H}_l^* + N_r \widehat{H}_r^*\right). \tag{34}$$

Again considering the binaural signal model (3) and substituting (34) back into (31), the target signal will be canceled out, and the error signals will consists of only the filtered noise components as follows:

$$\begin{aligned} E_l &= e^{-j\frac{2\pi}{M}\lambda L} \left(N_l \left(1 - \left|\widehat{H}_l\right|^2\right) - N_r \widehat{H}_l \widehat{H}_r^*\right), \\ E_r &= e^{-j\frac{2\pi}{M}\lambda L} \left(N_r \left(1 - \left|\widehat{H}_r\right|^2\right) - N_l \widehat{H}_r \widehat{H}_l^*\right). \end{aligned} \tag{35}$$

By transforming (35) into the PSD domain, we have

$$\widehat{\Phi}_e = \mathbf{A} \Phi_n - 2\mathrm{Re}\left\{\widehat{H}_l^* \widehat{H}_r\right\} \Phi_{n_l n_r} \widehat{\mathbf{H}}', \tag{36}$$

where $\widehat{\Phi}_e = \left[\widehat{\Phi}_{e_l}\ \widehat{\Phi}_{e_r}\right]^T$ and $\Phi_n = \left[\Phi_{n_l n_l}\ \Phi_{n_r n_r}\right]^T$ are a concatenation of the left and right error and noise PSDs, respectively. The matrix $\mathbf{A}$ is defined as

$$\mathbf{A} = \begin{bmatrix} \left(1 - \left|\widehat{H}_l\right|^2\right)^2 & \left|\widehat{H}_l\right|^2 \left|\widehat{H}_r\right|^2 \\ \left|\widehat{H}_l\right|^2 \left|\widehat{H}_r\right|^2 & \left(1 - \left|\widehat{H}_r\right|^2\right)^2 \end{bmatrix}, \tag{37}$$

while $\widehat{\mathbf{H}}' = \left[1 - \left|\widehat{H}_l\right|^2\ \ 1 - \left|\widehat{H}_r\right|^2\right]^T$.

Due to the bin-wise norm normalization, $\det(\mathbf{A})$ is very small, and thus, $\mathbf{A}$ is singular, regardless of the position of the target speaker. To solve the rank deficiency of $\mathbf{A}$, the noise PSDs at the left and right ear are again assumed to be identical, i.e., $\Phi_{n_l n_l} = \Phi_{n_r n_r} = \Phi_n$. Therefore, (36) is rewritten as

$$\widehat{\Phi}_e = \mathbf{B} \Phi_n - 2\mathrm{Re}\{\widehat{H}_l^* \widehat{H}_r\} \Phi_{n_l n_r} \widehat{\mathbf{H}}', \tag{38}$$

with

$$\mathbf{B} = \begin{bmatrix} \left|\widehat{H}_l\right|^2 \left|\widehat{H}_r\right|^2 + \left(1 - \left|\widehat{H}_l\right|^2\right)^2 \\ \left|\widehat{H}_l\right|^2 \left|\widehat{H}_r\right|^2 + \left(1 - \left|\widehat{H}_r\right|^2\right)^2 \end{bmatrix}. \tag{39}$$

4.3.1 Uncorrelated noise

Assuming uncorrelated noise, i.e., $\Phi_{n_l n_r} = 0$, (38) will be simplified to

$$\widehat{\Phi}_e(\kappa, \lambda) = \mathbf{B}(\kappa, \lambda) \Phi_n(\kappa, \lambda), \tag{40}$$

which is an over-determined problem. Thus, (40) can be solved using least-squares [55],

$$\widehat{\Phi}_n = \left(\mathbf{B}^T\mathbf{B}\right)\mathbf{B}^T\widehat{\Phi}_e. \tag{41}$$

Many practical noise situations, however, have to be modeled as diffuse noise [22], with high correlation in the low frequencies. Therefore, the noise PSD is underestimated especially at low frequencies.

4.3.2 Diffuse noise

Assuming an isotropic homogeneous noise field, the noise will be correlated in low frequencies and uncorrelated in high frequencies. Under the assumption of equal noise PSD at the left and right ear and substituting $\Phi_{n_l n_r} = \Phi_{n_r n_l} = \Gamma_{n_l n_r}\Phi_n$ into (38), we have

$$\begin{aligned}\Phi_e(\kappa,\lambda) = \mathbf{B}(\kappa,\lambda)\Phi_n(\kappa,\lambda) \\ - 2\mathrm{Re}\left\{\widehat{H}_l^*\widehat{H}_r\right\}\Phi_n\Gamma_{n_l n_r}\widehat{\mathbf{H}}'.\end{aligned} \tag{42}$$

The noise PSD then again can be estimated by solving (42) in the least-squares sense [55] as

$$\widehat{\Phi}_n = \left(\mathbf{C}^T\mathbf{C}\right)\mathbf{C}^T\widehat{\Phi}_e, \tag{43}$$

where

$$\mathbf{C} = \mathbf{B} - 2\mathrm{Re}\left\{\widehat{H}_l^*\widehat{H}_r\right\}\Gamma_{n_l n_r}\widehat{\mathbf{H}}'. \tag{44}$$

5 Instrumental measures related to adaptive speech blocking

In this section, we will introduce and discuss the evaluation tools utilized in this contribution.

5.1 Speech leakage ratio (SLR)

The performance of the described speech blocking-based noise PSD estimators and, consequently, of the noise reduction algorithms depends on the target speech cancelation ability. The Hagerman method [56] is thus employed to calculate an SLR, $i \in \{l, r\}$,

$$\mathrm{SLR}_i = \frac{1}{Ml_t}\sum_{\lambda=1}^{M}\sum_{\kappa=1}^{l_t} 10\log_{10}\left(\frac{\widehat{\Phi}_{\tilde{e}_i}}{\widehat{\Phi}_{\tilde{y}_i}}\right), \tag{45}$$

with $\widehat{\Phi}_{\tilde{e}_i}$ being the PSD of $\tilde{e}_i = e_{i,-} + e_{i,+}$, where the signal $e_{i,+}$ is the blocking output when the noisy signal is utilized as an input, i.e., $y_{i,+}(k) = x_i(k) + n_i(k)$, and $e_{i,-}$ is the blocking output when the input signal is composed as $y_{i,-}(k) = x_i(k) - n_i(k)$. Similarly, $\widehat{\Phi}_{\tilde{y}_i}$ is computed as the PSD of $\tilde{y}_i = y_{i,-} + y_{i,+}$ The total number of frames for averaging is given as l_t. Thus, $\widehat{\Phi}_{\tilde{e}_i}$ can be considered as the speech leakage PSD, while $\widehat{\Phi}_{\tilde{y}_i}$ denotes the PSD of the direct speech signal. This method is well known for the separate evaluation of noise and speech components. Lower SLR is better. More information can be found in [56, 57].

5.2 Noise PSD ratio (LogErr measure)

The efficacy of speech enhancement algorithms highly depends on the accurate estimation of the noise PSD. Thus, we have employed an intermediate measure for evaluating the performance of the noise PSD estimators, $i \in \{l, r\}$,

$$\mathrm{LogErr}_i = \frac{1}{Ml_t}\sum_{\kappa=1}^{l_t}\sum_{\lambda=1}^{M}\left|10\log_{10}\frac{\Phi_{n_i n_i}(\lambda,\kappa)}{\widehat{\Phi}_{n_i n_i}(\lambda,\kappa)}\right|, \tag{46}$$

where $\Phi_{n_i n_i}$ and $\widehat{\Phi}_{n_i n_i}$ are the true and estimated noise PSDs. The true noise PSD is obtained according to (5), therein employing the given true effective noise signals, since algorithms based on short filters will attempt to estimate the effective noise.

6 Instrumental evaluation results

6.1 Experimental setup

The experiments are performed with the BRIRs *measured* in a reverberant "stairway" (direct-to-reverberation ratio, DRR = 11 dB), taken from the Aachen room impulse response database [58, 59], with a length of 5000 samples at a sampling frequency of $f_s = 16$ kHz. The location of the desired speaker can be between $-90° \le \theta \le 90°$, as illustrated in Fig. 1.

The left and right microphone signals are then generated by convolving the target speech signal with the binaural impulse responses. The clean speech signal is a 60-s concatenation of the female and male sentences taken from the TIMIT database [60]. A total of 30% of the total length consists of speech pause. Moreover, no initial noise-only frames have been utilized. Regarding the additive noise, six different binaural noises, including cafeteria noise, kindergarten noise, and Mensa noise, from the ETSI database [61] were used. Moreover, the computer-generated binaural babble noise and binaural white Gaussian noise (WGN) [62] were also considered in our evaluation.

It is furthermore instructive to investigate the performance of the proposed noise PSD estimators in the presence of nonstationary noise. This investigation addresses the capability of the proposed algorithm to track the noise PSD. To provide a reliable comparison, a modulated diffuse noise with different modulation frequencies with $i \in \{l, r\}$,

$$n_i(t) = n_{0,i}(t)(1 + 0.8\sin(2\pi f_m t/f_s)), \tag{47}$$

is considered as a reproducible dynamic noise model, where f_m is the modulation frequency varying from 0.05 to 1 Hz. The $n_{0,i}(t)$ is a computer-generated diffuse WGN [62] such that its coherence function follows a 2D head-related coherence model [51].

6.1.1 Algorithm parameters

All considered signals are sampled at $f_s = 16$ kHz and are segmented into 50% overlapping frames of length $M = 512$. The overlapping frames are then windowed using a square-root Hanning window and transformed into the frequency domain via the STFT of length M [46]. The smoothing factor for estimating the (cross-) power spectral densities is set to $\alpha = 0.8$ if not stated otherwise. The spectral correction gains are floored to -20 dB. The causality delay τ_a is set to 30 samples. The length of the adaptive filters is $L = 256$, while the length of the RIRs is 5000 samples. The stepsize μ_0 of ITBF, PCAB, and CRB are set to 0.1, 0.2, and 0.1, respectively. Moreover, the over-subtraction factor β and the spectral flooring $G_{\min}$ of the cue-preserving MMSE gain function in (10) are set to 1.4 and -20 dB, respectively. The adaptive speech blocking filters are realized with the FDAF [49].

6.1.2 Selected algorithms for comparison

To investigate the performance of a wide range of subspace algorithms for noise PSD estimation, we compare the performance of the principal-component-analysis based estimator, i.e., (*PCAB*), with the noise PSD estimator based on the interaural transfer function (*ITFB*), [41], and with the noise PSD estimator relying on cross-relation error minimization (*CRB*) [42], therein considering the diffuse noise assumption.

Moreover, for the sake of completeness, the studied speech and noise-subspace noise PSD estimators are compared to other binaural and single-channel noise PSD estimators available in the literature: the improved CPSD method (*ImCPSD*) [22] and the single-channel SPP-based method (*SC-SPP*) [63]. It should be mentioned that [22] used the same error signals as described in (12). The noise PSDs estimated by the different algorithms are then utilized in the cue-preserving MMSE filter to deliver the enhanced microphone signals. The enhanced signal using a priori known "true" noise PSD is denoted as *Ref*.

6.2 Investigation of speech leakage

Due to the estimation error in the interaural and source-to-microphone transfer functions, for instance, due to noise or reverberation, the speech components will leak to some degree into the blocking residual. These leaked speech components hence result in noise power overestimation and consequently in speech distortion after the enhancement stage. Therefore, it is crucial for blocking-based noise PSD estimators to exhibit small speech leakage.

Figure 3 shows the computed SLR according to (45) as a function of input SNR for different blocking algorithms. Here, "*CS*" denotes the input clean speech signal power. It can be clearly observed that all algorithms in all SNRs under consideration generally attenuate the input speech power. The CRB achieves the lowest SLR. This is because for CRB, the effective error of the channel identification can be appropriately approximated by a common transfer function. The SLR in ITFB at low SNR is large because ITFB faces greater difficulties in the unbiased estimation of interaural transfer functions. This is because the respective Wiener solution of the filter is biased by the noise PSD [30]. Due to the inverse filtering problem in ITFB, the SLR cannot be reduced even at high SNRs.

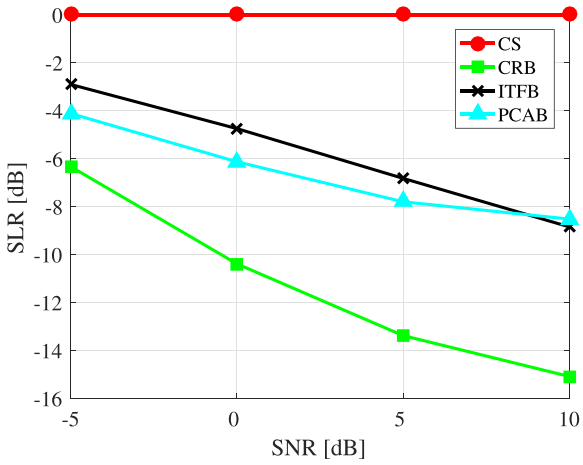

Fig. 3 SLR as a function of SNR for different algorithms, $\theta = -45°$

To better elaborate upon the differences in the studied algorithms in the blocking of the speech components, Fig. 4 illustrates the SLR at different azimuth angles. Here, apparently, the lowest SLR can be found in the frontal direction. Moreover, the ranking of the speech and noise-subspace noise PSD estimators in the residual speech attenuation is preserved in comparison to Fig. 3.

The performance of the blocking systems can be evaluated additionally in terms of system identification. In this

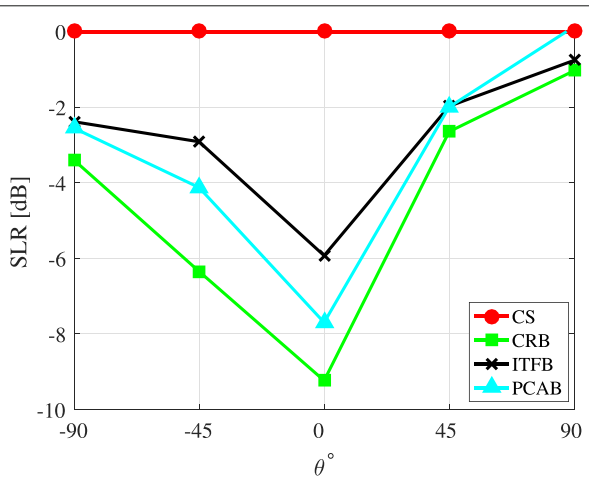

Fig. 4 Computed SLR as a function of azimuth angle for different speech blocking algorithms, where SNR = −5 dB

study, we have chosen not to present the related results because the system identification problem in the presence of ambient noise has been widely studied in the literature. For instance, for ITFB, refer to [30]; for CRB, see [44, 52]; and for PCAB, more information can be found in [44, 54]. The results discussed in the aforementioned studies are confirmed by our investigations.

6.3 Noise PSD estimation

To evaluate the performance of the studied algorithms in highly non-stationary noise environments, the binaural modulated babble noise (47) with different modulation frequencies is employed. Figure 5 shows the computed LogErr as a function of the modulation frequency for different algorithms. All the blocking-based noise PSD estimators are apparently extremely robust against dynamic noise conditions, in sharp contrast to SC-SPP. The estimated noise PSD for the modulation frequency of 1 Hz is illustrated in Fig. 6 as a function of time. It can be confirmed here that the blocking-based noise PSD estimators are able to track the noise power changes quickly, whereas the SC-SPP cannot follow the time-varying noise PSD properly.

Because more realistic noisy conditions are of great interest in audio signal processing, the LogErr measure, averaged over different realistic noise types, is presented in Fig. 7. As can be observed from the experimental results, all blocking-based algorithms yield smaller LogErr in comparison to the SC-SPP. Among the blocking-based estimators, ITFB is superior because it provides binaural noise PSD estimation. It is followed by the ImCPSD algorithm [22], which employs a similar error signal as described in (12).

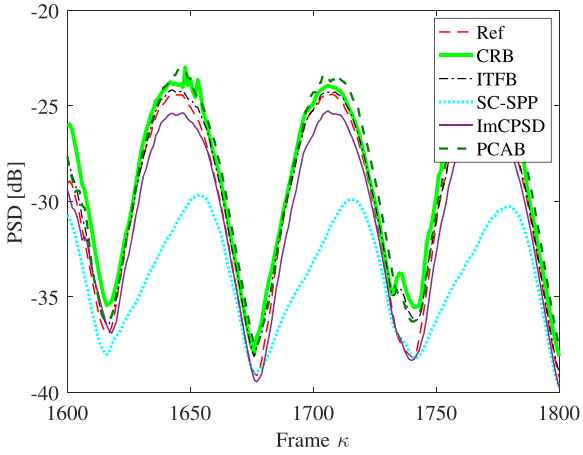

Fig. 6 Comparison of estimated noise PSD as a factor of time at the *left ear* by different algorithms. Here, $f_m = 1$ Hz and SNR $= -5$ dB, $\theta = -90°$, averaged over all frequencies

6.4 Noise reduction

The segmental SNR improvement [38] and the perceptual evaluation of speech quality (PESQ) [64] are used to assess the overall speech enhancement performance of the algorithm. The cue-preserving MMSE filter (10) is computed using the estimated PSDs. For a fair comparison, the smoothing factor in the PSD estimator was set $\alpha = 0.8$ in all algorithms where was needed. All results are *averaged* across the left and right ears and across all considered noise types.

The results of the segmental SNR improvement in Fig. 8a shows that the speech-blocking-based algorithms obtain better improvements in segmental SNR at almost all SNRs in comparison to the other studied algorithms.

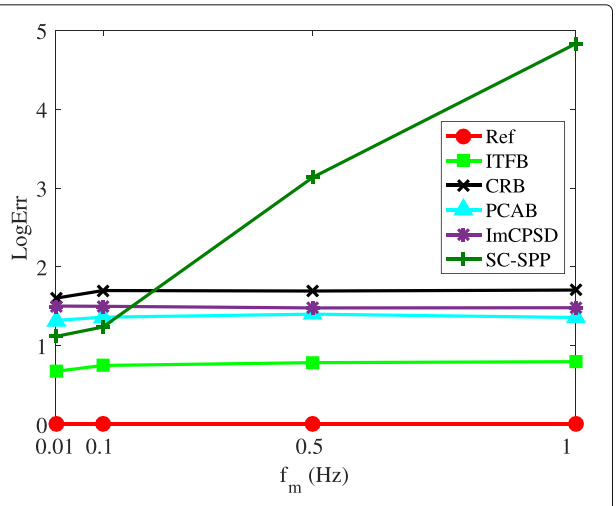

Fig. 5 Comparison of LogErr averaged over all frame indices and frequency bins as a function of the modulation frequency of dynamic noise. SNR $= -5$ dB, $\theta = -90°$

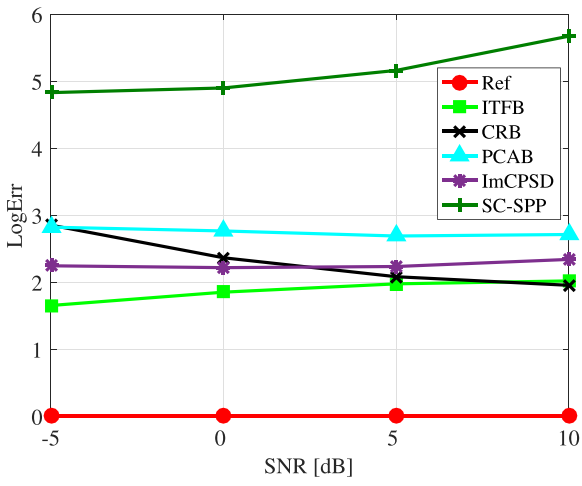

Fig. 7 Comparison of LogErr according to (46) averaged over all frame indices and frequency bins with different noise types and input SNRs, where $\theta = -90°$

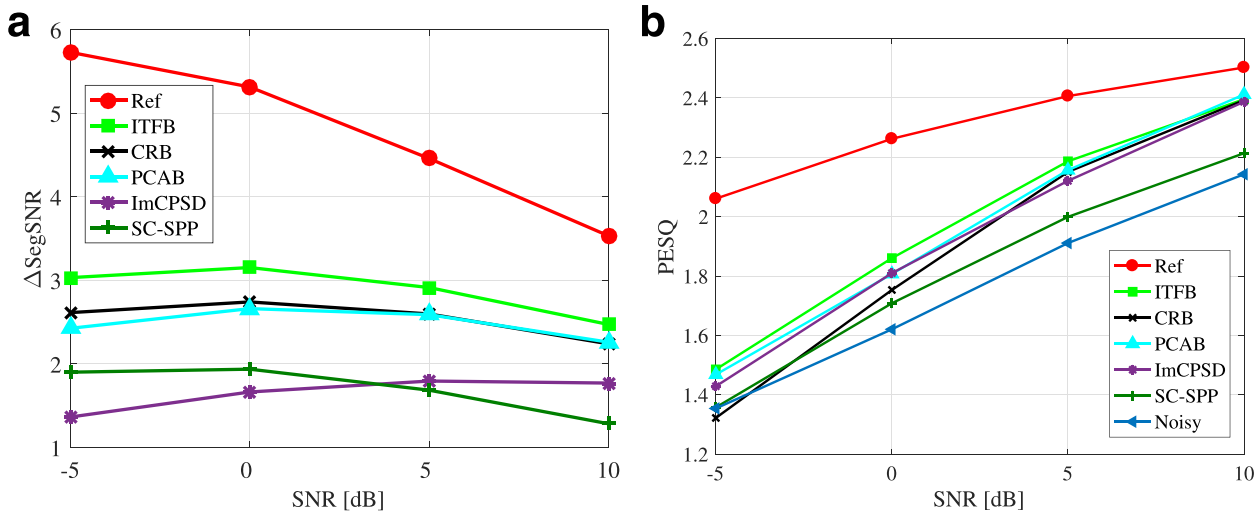

Fig. 8 Speech enhancement comparison in terms of **a** ΔSegSNR and **b** PESQ for different algorithms averaged over all noise types ($\theta = -90°$)

The ITFB achieves a superior noise suppression performance because it provides binaural noise estimates as well as a small error in the LogErr, as previously shown in Figs. 7 and 5.

The results of the PESQ measure are presented in Fig. 8b. Similarly, we can see that the ITFB and PCAB improve the PESQ score at all SNRs. At high SNR, e.g., SNR = 10 dB, all the studied algorithms could achieve improved PESQ scores, except for SC-SPP. However, the differences in the PESQ scores between the considered algorithms are small and not one-to-one related to the LogErr results, as shown in Fig. 7. The spectral flooring in the cue-preserving MMSE gain (10), for instance, reduces the influence of the estimated noise PSD on the PESQ score. The results from all measures under consideration are slightly different because each measure illustrates specific characteristics of the signal.

The remaining gap between the best performing algorithm and the "Ref", i.e., given the true noise PSD, can be explained by the fact that there is no speech leakage involved in the true noise PSD. Moreover, the reference case employs the true binaural noise PSD in the left and right ear, which is of particular importance in non-stationary noise frames. In other words, the aforementioned gap can be reduced by employing precisely estimated noise PSDs at the left and right ears and by further reducing the speech leakage in the blocking residual.

6.5 Binaural cue preservation

Binaural cue preservation is one of the main quality factors that need to be considered in addition to noise reduction and speech preservation in binaural speech enhancement. Preserving the binaural cues of the speech signal, particularly ILD and ITD, helps the listener to localize the desired speaker more precisely.

The bilateral gain functions $G_i = 1 - \frac{\phi_{n_i n_i}}{\phi_{y_i y_i}}$ with $i \in \{l, r\}$ and the binaural cue-preserving MMSE filter in (10) are compared in terms of binaural cue preservation. Here, the ILD and ITD are estimated according to [65] using the shadow-filtered clean signal. It should be noted that only frequency ranges higher than 1.5 kHz and lower than 1.5 kHz are considered for the computation of the ILD and ITD, respectively. The ambient noise is the isotropic diffuse noise generated by the algorithm in [62] with the 2D coherence model at 0 dB SNR.

The ΔILD and ΔITD are the deviations of the processed ILD and ITD by the binaural and bilateral gain functions from the ITD and ILD of the input clean speech signal in each frequency and frame, respectively. The averaged ΔILD and ΔITD over the frames and frequencies are then reported in Fig. 9. As shown, the corresponding errors in both the ILD and ITD are higher for the conventional bilateral gain functions, while the cue-preserving MMSE filter keeps the binaural cues undistorted. The proposed binaural cue-preserving MMSE filter preserves the binaural cues with a slight loss in the noise reduction performance. This is depicted in Fig. 10, where the true noise PSDs are utilized. The noise reduction performance degradation will be negligible when the estimated noise PSDs are used (not shown here).

7 Subjective evaluation

A subjective listening test is the most appropriate way to assess the effect of the speech enhancement algorithms [66–69]. Thus, various methods and procedures have been used and developed, for instance, for the assessment

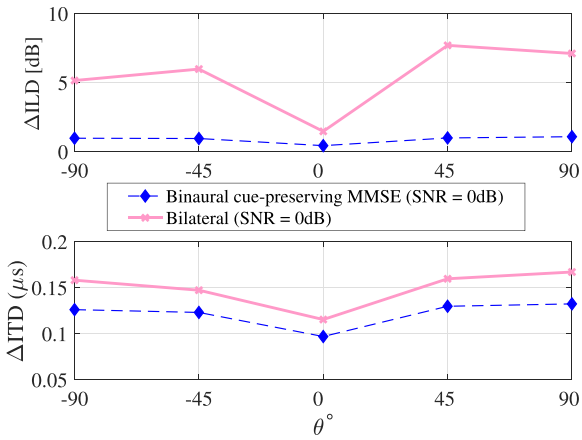

Fig. 9 Upper: ΔILD and bottom: ΔITD correspond to the binaural and bilateral gain functions, (SNR = 0 dB)

of the speech quality [70, 71], speech intelligibility [72], and spatial cue preservation [73].

In this contribution, we also developed a listening test based on a real-time signal-processing platform to evaluate the robustness and validity of the algorithms in a realistic setting. However, the employed overlap-add framework in the algorithm design and the utilized USB sound card in the demonstration setup do not allow for very small latencies for sound processing. Therefore, the real-time listening processing here mainly implies the online execution of the adaptive algorithms.

Because the employed real-time listening test is a new procedure and because the exact form of the test for the evaluation of the noise reduction algorithms is not yet available, we developed a test procedure according to the perceptual evaluation preparation process suggested in [74]. However, the standardized methods recommended in [75–77] are accommodated in different stages of the listening test, as we wish to rely on proven methods as much as possible.

The algorithms are implemented on a single-board Raspberry Pi computer [78]. The implementation of the considered algorithms is realized in Simulink [79], a graphical programming and development environment. Using a complementary support package for Simulink, the Raspberry Pi is conveniently interfaced. Because the proposed solutions suppress the noise without any assumption on the noise PSD, target speaker location or voice activity detection (VAD), they can be conveniently evaluated and compared in real time [80].

7.1 Experimental setup

The experiment is conducted in a medium-sized room with a reverberation time of approximately 200 ms at the Institute of Communication Acoustics, Bochum. The room schematic and the experimental setup are illustrated in Fig. 11. The target speaker (operator) walks in front of the Head-and-Torso Simulator (HATS) while speaking. The path along which the operator (speaker) mostly walks is also depicted in Fig. 11 as a dashed line. The distance between the operator and the HATS is approximately 70 cm. The speech material consists of the natural speech of the authors (female/male) for on the order of 15 min per subject. The target speech is superimposed with an approximation of ideally diffuse background noise at an SNR of approximately 6 dB according to the Lombard effect. As shown in Fig. 11, four loudspeakers play four independent babble noise signals to generate a diffuse noise. The individual loudspeakers were calibrated to deliver an equal noise level at the location of the HATS. The microphones embedded in the Sony MDR-NC31EM headset capture the noisy signals at the left and right ears of the HATS. The captured noisy signals are then fed into

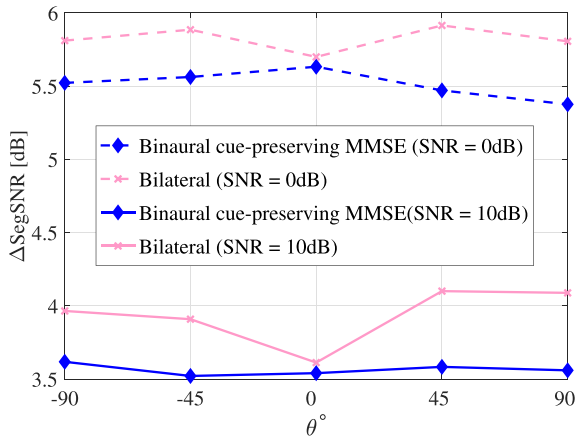

Fig. 10 Comparison of the proposed binaural cue-preserving MMSE filter and the bilateral filter in segmental SNR improvement

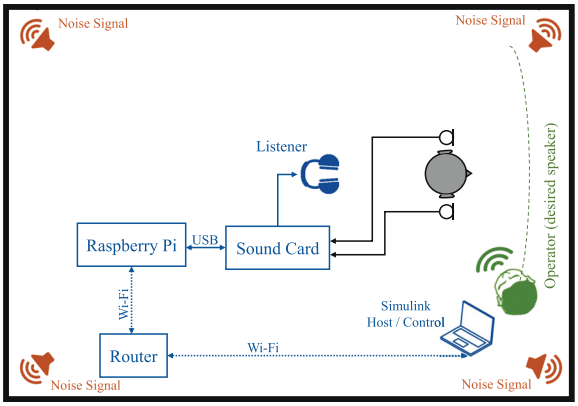

Fig. 11 Schematic of the demonstration/evaluation room

the Focusrite Scarlett 2i2 external USB sound card and transferred to the Raspberry Pi for real-time processing.

The processed signals are presented to the subjects (listeners) over a passive sound-isolated headphone (Sennheiser HDA200) at a sound level that the subjects find convenient, approximately 70 dB SPL when the noise level is 65 dB SPL. As shown in Fig. 11, the host computer offers the operator the possibility to alternately provide the listener with the processed signal by different binaural noise reduction algorithms, including ITFB, CRB, and PCAB, in addition to the unprocessed signal.

7.2 Subjective listening test

A total of 14 normal-hearing subjects, including 11 males and 3 females, from 25 to 40 years old, participated in this real-time assessment of the binaural noise reduction algorithms. Although the normal-hearing people and the hearing-aid users would perceive the enhanced sound quality differently [81], in this work, we only rely on normal-hearing subjects. The participants were asked to sit right behind and close to the HATS, keeping the direction of their head and of their body similar to the that of the HATS if possible (Fig. 11).

To simulate a realistic noisy condition that occurs in daily life, a conversational test [82] has been employed here. A scientific discussion is conducted between the operator (speaker) and the listener, who wears the headphones during the conversation and hence is virtually in the position of the HATS. However, due to the effect of the delayed auditory feedback [83], which makes the listener hear his/her own voice, the conversation is mostly one-sided. The stimulus is the operator speech signal superimposed with the diffuse noise. The location of the operator is varied to evaluate the robustness of the studied algorithms to time-varying BRIRs and hence varying binaural cues.

The processed sounds are presented to the listener by switching among the considered noise reduction algorithms. In the training phase, the listener has to listen to all processed signals at least once to appreciate the context of the presented audio signals. Following the training phase, the evaluation stage is started, and the listener is asked to evaluate the signals on a continuous scale between 0 and 100 [66]. The audio signals are presented to the listeners repeatedly if requested. A more detailed specification of the available scores employed in this listening test is presented in Table 1.

7.3 Investigated attributes

The subjects were all expert listeners, and the training phase was conducted by briefing the listeners on the meanings and possible impairments of the attributes in the processed signal. The studied quality attributes, together with possible impairments per attribute, are summarized in Table 2. The listeners received an instruction sheet including the score specifications (Table 1), and the attribute definitions with related impairments (Table 2) before the test was started. Moreover, the clean speech was presented to the subjects in the briefing session as a part of the training phase before the test began. This was done to equalize the subjects' opinions on the perceived quality with respect to the available attributes and the rating scale as much as possible. The listening test was found to be a very realistic representation of a daily noisy situation by the operator and listeners.

Table 1 Score specification for rating the binaural noise reduction algorithms

Score		Explanation
80–100	Excellent	No impairment is audible, great performance
60–80	Good	Nice utility that mostly meets expectations
40–60	Fair	Mostly acceptable, but some undesired impairments are already detected
20–40	Poor	Presence of harsh impairments that leave no doubt of insufficiency
0–20	Bad	No utility

The hypothesis that the listening test results follow a normal distribution is rejected by the Anderson-Darling test [84]. Because the data were not normally distributed, we used the Kruskal-Wallis test [85] for variance analysis. We compared the performance of the algorithms with respect to each attribute. For example, for background noise attenuation, this comparison was meant to examine whether there were significant differences between different blocking-based algorithms and the noisy signal. For the speech quality assessment, the significant differences between the unprocessed and processed signals were not expected, as the speech signals should be kept undistorted through the processing.

The results of the listening test are summarized in Fig. 12, including the estimated median values indicated in

Table 2 Explanation of investigated attributes of the processed signal along with possible related impairments

Attribute	Meaning	Possible impairment
Speech quality	Utilitarian comparison of the speech quality w.r.t. the assumed original	Speech onset suppression, artificial reverberation, or speech spectral smearing
Background noise attenuation	Noise attenuation	Noise level was annoying or unacceptably loud
Residual noise naturalness	Naturalness of the residual noise	Musical noise perception
Speech spatial sound cues	Consistency of speech spatial cues w.r.t. simultaneous visual cues	Spatial desynchronization

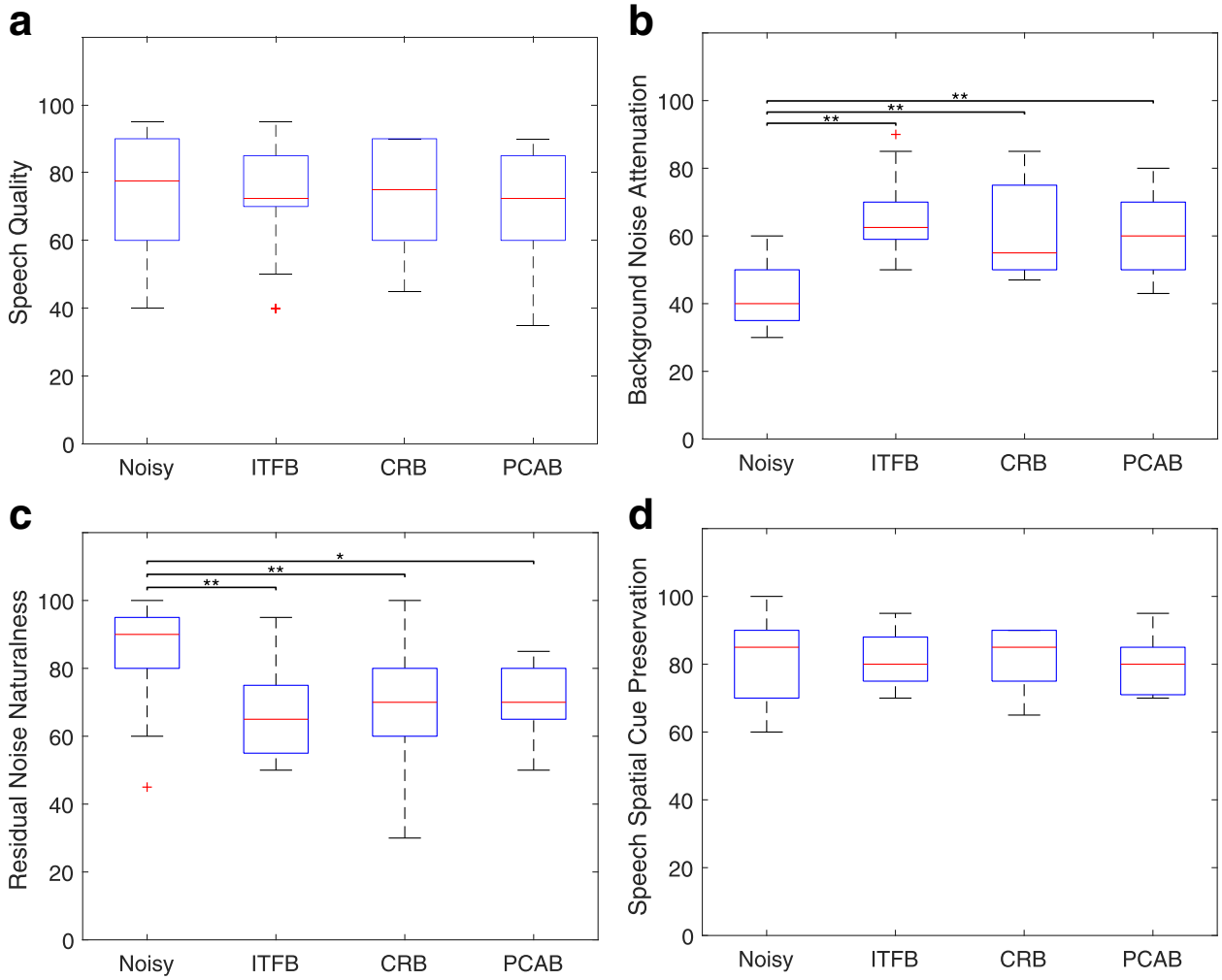

Fig. 12 Boxplots for **a** speech quality, **b** background noise attenuation, **c** residual noise naturalness, and **d** speech spatial cue preservation for an approximately diffuse noise, where SNR $\approx$ 6 dB according to the Lombard effect

red. The statistical significance of the medians is indicated by the square brackets on top of each boxplot. It should be noted that one asterisk corresponds to $p < 0.05$, while two asterisks represent $p < 0.01$.

It is observed from Fig. 12a that all algorithms achieved a very good perceived speech quality. Due to the high amount of ambient noise, the listener had difficulty focusing on the speech signal in the evaluation of the speech quality. Therefore, the variance is high in the speech quality of the unprocessed signal.

The comparison of algorithms in terms of background noise attenuation is presented in Fig. 12b. As can be seen, the listeners rated the processed signals as significantly superior to the noisy signal in terms of noise attenuation. The ITFB and PCAB were perceived to have performed similarly well in suppressing the background noise according to the median values.

In terms of the residual noise naturalness, presented in Fig. 12c, the unprocessed noise was rated significantly more natural in comparison to the processed noise by different algorithms. However, this is not surprising considering noise artifacts; for instance, musical noise is one of the well-known drawbacks of the Wiener-type noise reduction methods [30]. With respect to median values, the ITFB was perceived to be slightly more aggressive toward the noise signals, which can be additionally confirmed by the objective evaluation results presented in Fig. 8a.

The speech spatial cue rating is presented in Fig. 12d. As can be seen, the algorithms are rated similarly according to the median values, and there are no significant differences between the unprocessed and processed signals. The listeners rated the speech spatial cue preservation by how consistent they perceived the spatial cues with

respect to the visual cues. Because the listeners were wearing headphones at all times during the test, some of the listeners did not experience natural speech cue perception due to the use of the headphones. Therefore, there is a considerably high variance in all the signals.

8 Conclusions

In this contribution, a binaural cue-preserving gain function based on the MSE criterion is proposed for binaural noise reduction. A comparison of the proposed gain function and a bilateral Wiener filter has been conducted and shows that the binaural cues, particularly ILD and ITD, can be remarkably preserved by applying the proposed gain function without experiencing a considerable loss in noise reduction performance.

Moreover, a class of binaural noise PSD estimators based on speech blocking has been discussed. The noise PSD estimators rely on adaptive target speech cancelation. The comparison reveals individual strengths and weaknesses. For instance, ITFB provides binaural noise estimation, which is one of the key factors toward achieving a performance similar to the ideal reference noise reduction. The CRB, in turn, provides the lowest speech leakage, which is another key factor. These factors are in line with our observations from the real-time evaluation.

Furthermore, a real-time subjective listening test has been developed to assess the performance of blocking-based algorithms in a realistic acoustic environment. The listening test data analysis verifies the objective evaluation outcomes.

Acknowledgements
The authors acknowledge Prof. Rainer Martin for his valuable feedback.

Authors' contributions
All the contributions are by the authors. Both authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Received: 27 March 2017 Accepted: 15 June 2017
Published online: 10 July 2017

References

1. C Mathers, A Smith, M Concha, Global burden of hearing loss in the year 2000. World Health Organization (2000)
2. TVD Bogaert, TJ Klasen, M Moonen, LV Deun, J Wouters, Horizontal localization with bilateral hearing aids: without is better than with. J. Acoust. Soc. Am. **119**(1), 515–526 (2006)
3. S Doclo, R Dong, TJ Klasen, J Wouters, S Haykin, M Moonen, in *Proc. IEEE Intl. Workshop on Acoustic Echo and Noise Control (IWAENC)*. Extension of the multi-channel Wiener filter with localization cues for noise reduction in binaural hearing aids, (Eindhoven, 2005), pp. 221–224
4. Y Suzuki, S Tsukui, F Asano, R Nishimura, New design method of a binaural microphone array using multiple constraints. IEICE Trans. Fundamentals Electron. Commun. Comput. Sci. **82**(4), 588–596 (1999)
5. J Szurley, A Bertrand, BV Dijk, M Moonen, Binaural noise cue preservation in a binaural noise reduction system with a remote microphone signal. IEEE/ACM Trans. Audio, Speech Lang. Process. **24**(5), 952–966 (2016)
6. S Haykin, KJR Liu, in *Handbook on Array Processing and Sensor Networks*, ed. by S. Doclo, MMS Gannot, A Spriet. Acoustic beamforming for hearing aid applications (Wiley, New York, 2008), pp. 269–302
7. B Cornelis, S Doclo, TV den Bogaert, M Moonen, J Wouters, Theoretical analysis of binaural multimicrophone noise reduction techniques. IEEE Trans. Audio, Speech, Lang. Process. **18**(2), 342–355 (2010)
8. S Doclo, TJ Klasen, TV den Bogaert, J Wouters, M Moonen, in *Proc. Int. Workshop Acoustic Echo Noise Control (IWAENC)*. Theoretical analysis of binaural cue preservation using multi-channel Wiener filtering and interaural transfer functions, (Paris, 2006)
9. M Azarpour, G Enzner, R Martin, in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*. Adaptive binaural noise reduction based on matched-filter equalization and post-filtering, (Vancouver, 2013), pp. 1–4
10. E Hadad, D Marquardt, S Doclo, S Gannot, Theoretical analysis of binaural transfer function MVDR beamformers with interference cue preservation constraints. IEEE Trans. Audio, Speech, Lang. Process. **23**(12), 2449–2464 (2015)
11. MH Costa, PA Naylor, in *in Proc. IEEE Signal Processing Conf. (EUSIPCO)*. ILD preservation in the multichannel Wiener filter for binaural hearing aid applications, (Lisbon, 2014)
12. TJ Klasen, TV den Bogaert, M Moonen, J Wouters, Binaural noise reduction algorithms for hearing aids that preserve interaural time delay cues. IEEE Trans. Signal Process. **55**(4), 1579–1585 (2007)
13. TV 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)
14. TVD Bogaert, J Wouters, S Doclo, M Moonen, in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. Binaural cue preservation for hearing aids using an interaural transfer function multichannel Wiener filter, vol. 4, (Honolulu, 2007), pp. 565–568
15. E Hadad, S Doclo, S Gannot, The binaural LCMV beamformer and its performance analysis. IEEE/ACM Trans. Audio, Speech, Lang. Process. **24**(3), 543–558 (2016)
16. D Marquardt, E Hadad, S Gannot, S Doclo, Theoretical analysis of linearly constrained multi-channel Wiener filtering algorithms for combined noise reduction and binaural cue preservation in binaural hearing aids. IEEE Trans. Audio, Speech, Lang. Process. **23**(12), 2384–2397 (2015)
17. D Marquardt, V Hohmann, S Doclo, in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. Binaural cue preservation for hearing aids using multi-channel Wiener filter with instantaneous ITF preservation, (Kyoto, 2012), pp. 21–24
18. D Marquardt, V Hohmann, S Doclo, in *2014 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. Perceptually motivated coherence preservation in multi-channel Wiener filtering based noise reduction for binaural hearing aids, (Florence, 2014), pp. 3660–3664
19. D Marquardt, V Hohmann, S Doclo, in *IEEE Int. Conf. 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
20. D Marquardt, V Hohmann, S Doclo, in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. Interaural coherence preservation in MWF-based binaural noise reduction algorithms using partial noise estimation, (Brisbane, 2015), pp. 654–658
21. D Marquardt, V Hohmann, S Doclo, Interaural coherence preservation in multi-channel Wiener filtering-based noise reduction for binaural hearing aids. IEEE Trans. Audio, Speech, Lang. Process. **23**(12), 2162–2176 (2015)
22. AH Kamkar-Parsi, M Bouchard, Improved noise power spectrum density estimation for binaural hearing aids operating in a diffuse noise field environment. IEEE Trans. Audio, Speech, Lang. Process. **17**(4), 521–533 (2009)
23. N Yousefian, JHL Hansen, PC Loizou, A hybrid coherence model for noise reduction in reverberant environments. IEEE Signal Process. Lett. **22**(3), 279–282 (2015)
24. M Jeub, M Schäfer, T Esch, P Vary, Model-based dereverberation preserving binaural cues. IEEE Trans. on Audio, Speech, Lang. Process. **18**, 1732–1745 (2010)

25. F Mustière, M Bouchard, H Najaf-Zadeh, R Pichevar, L Thibault, H Saruwatari, Design of multichannel frequency domain statistical-based enhancement systems preserving spatial cues via spectral distances minimization. Signal Process. Elsevier. **93**(1), 321–325 (2013)
26. A Tsilfidis, E Georganti, J Mourjopoulos, in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. Binaural extension and performance of single-channel spectral subtraction dereverberation algorithms, (Prague, 2011), pp. 1737–1740
27. B Kollmeier, J Peissig, V Hohmann, Real-time multiband dynamic compression and noise reduction for binaural hearing aids. J. Rehab. Res. Dev. **30**(1), 82–94 (1993)
28. M Dörbecker, S Ernst, in *Proc. of European Signal Processing Conf. (EUSIPCO)*. Combination of two-channel spectral subtraction and adaptive Wiener post-filtering for noise reduction and dereverberation, (Trieste, 1996), pp. 995–998
29. AH Kamkar-Parsi, M Bouchard, Instantaneous binaural target PSD estimation for hearing aid noise reduction in complex acoustic environments. IEEE Trans. Instrumentation Meas. **60**(4), 1141–1154 (2011)
30. P Vary, R Martin, *Digital Speech Transmission. Enhancement, Coding and Error Concealment*. (John Wiley & Sons , Ltd, Chichester, 2006)
31. N Wiener, *Extrapolation, Interpolation and Smoothing of Stationary Time Series*. (John Wiley & Sons, New York, USA, 1949)
32. JS Lim, AV Oppenheim, Enhancement and bandwidth compression of noisy speech. Proc. IEEE. **67**(12), 1586–1604 (1979)
33. JHL Hansen, MA Clements, Constrained iterative speech enhancement with application to speech recognition. IEEE Trans. Signal Process. **39**(4), 795–805 (1991)
34. Y Ephraim, D Malah, Speech enhancement using a minimum meansquare error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech, Signal Process. **32**(6), 1109–1121 (1984)
35. T Lotter, P Vary, Dual-channel speech enhancement by superdirective beamforming. EURASIP J. Adv. Signal Process. **2006**, 1–14 (2006)
36. R Zelinski, in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. A microphone array with adaptive post-filtering for noise reduction in reverberant rooms, vol. 5, (New York, 1988), pp. 2578–2581
37. IA McCowan, H Bourlard, Microphone array post-filter based on noise field coherence. IEEE Trans. Speech Audio Process. **11**(6), 709–716 (2003)
38. PC Loizou, *Speech Enhancement: Theory and Practice*, 1st edn. (CRC Press, Inc., Florida, 2007)
39. L Wang, T Gerkmann, S Doclo, in *Proc. Int. Workshop on Acoustic Signal Enhancement (IWAENC)*. Noise PSD estimation using blind source separation in a diffuse noise field, (Aachen, 2012), pp. 1–4
40. K Reindl, Y Zheng, A Schwarz, S Meier, R Maas, A Sehr, W Kellermann, A stereophonic acoustic signal extraction scheme for noisy and reverberant environments. Comput. Speech Lang. **27**(3), 726–745 (2013)
41. M Azarpour, G Enzner, R Martin, in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*. Binaural noise PSD estimation for binaural speech enhancement, (Florence, 2014)
42. M Azarpour, G Enzner, in *Int. Workshop on Acoustic Signal Enhancement (IWAENC)*. Fast noise PSD estimation based on blind channel identification, (Antibes Juan les Pins, French Riviera, 2014), pp. 223–227
43. A Hyvärinen, J Karhunen, E Oja, *Principal Component Analysis*. (John Wiley & Sons, New York, 2001)
44. G Enzner, I Merks, T Zhang, in *Proc. of the 20th European Signal Processing Conf. (EUSIPCO)*. Adaptive filter algorithms and misalignment criteria for blind binaural channel identification in hearing-aids, (Bucharest, 2012), pp. 315–319
45. JC Junqua, The Lombard reflex and its role on human listeners and automatic speech recognizers. J. Acoust. Soc. Am. **93**(1), 510–524 (1993)
46. JB Allen, Short term spectral analysis, synthesis, and modification by discrete Fourier transform. IEEE Trans. Acoust Speech, Signal Process. **25**(3), 235–238 (1977)
47. AV Oppenheim, RW Schafer, *Discrete-Time Signal Processing*. (Prentice Hall, Englewood Cliffs, 1989)
48. G Enzner, JSM Azarpour, in *Proc. Int. Workshop on Acoustic Signal Enhancement (IWAENC)*. Cue-preserving MMSE filter for binaural speech enhancement, (2016)
49. S Haykin, *Adaptive Filter Theory*. (Prentice Hall, Upper Saddle River, New Jersey, New Jersy, 2001)
50. H Kuttruff, *Room Acoustics*, 5th edn. (Spon Press, Abingdon, 2009)
51. M Jeub, M Dorbecker, P Vary, A semi-analytical model for the binaural coherence of noise fields. IEEE Signal Process. Lett. **18**(3), 197–200 (2011)
52. D Schmid, G Enzner, Cross-relation-based blind SIMO identifiability in the presence of near-common zeros and noise. IEEE Trans. Signal Process. **60**(1), 60–72 (2012)
53. J Benesty, MM Sondhi, YA Huang (eds.), *Springer Handbook of Speech Processing* (Springer, Berlin Heidelberg, 2008)
54. E Warsitz, R Haeb-Umbach, Blind acoustic beamforming based on generalized eigenvalue decomposition. IEEE Trans. Audio, Speech, Lang. Process. **15**(5), 1529–1539 (2007)
55. JH Wilkinson, C Reinsch, *Linear Algebra*. (Springer, Berlin Heidelberg, 1971)
56. B Hagerman, A Olofsson, Nästén: Noise reduction measurements in hearing aids. Presentation at IHCON (2001)
57. H Björn, O Åke, A method to measure the effect of noise reduction algorithms using simultaneous speech and noise. Acta Acust United Ac. **90**, 356–361 (2004)
58. M Jeub, M Schäfer, P Vary, in *Proc. of Int. Conf. on Digital Signal Processing (DSP)*. A binaural room impulse response database for the evaluation of dereverberation algorithms, (Santorini, 2009), pp. 1–4
59. M Jeub, M Schäfer, H Krüger, CM Nelke, C Beaugeant, P Vary, in *Int. Congress on Acoustics (ICA)*. Do we need dereverberation for hand-held telephony? (Sydney, 2010), pp. 1–7
60. JS Garofolo, LF Lamel, WM Fisher, JG Fiscus, DS Pallett, NL Dahlgren, *DARPA TIMIT Acoustic-phonetic continuous speech corpus CDROM*. (NIST, 1993). http://www.ldc.upenn.edu/Catalog/LDC93S1.html
61. ETSI EG 202 396-1: Speech quality performance in the presence of background noise; Part 1: Background noise simulation technique and background noise database (2009)
62. EAP Habets, I Cohen, S Gannot, Generating nonstationary multisensor signals under a spatial coherence constraint. J. Acoustic Soc. Am. **124**(5), 2911–2917 (2008)
63. T Gerkmann, RC Hendriks, in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. Noise power estimation based on the probability of speech presence, (New Paltz, 2011), pp. 145–148
64. AW Rix, JG Beerends, MP Hollier, AP Hekstra, in *Proc. IEEE Int. Conf. Acoustic , Speech, Signal Processing (ICASSP)*. Perceptual evaluation of speech quality (PESQ)— a new method for speech quality assessment of telephone networks and codecs, vol. 2, (Salt Lake City, 2001), pp. 749–752
65. T May, S van de Par, A Kohlrausch, A probabilistic model for robust localization based on a binaural auditory front-end. IEEE Trans. Audio, Speech Lang. Process. **19**(1), 1–13 (2011)
66. S Bech, N Zacharov (eds.), *Perceptual Audio Evaluation—Theory, Method and Application* (John Wiley & Sons, Chichester, England, 2006)
67. E Parizet, VN Nosulenko, Multi-dimensional listening test: selection of sound descriptors and design of the experiment. Noise Control Eng. J. **47**(6), 1–6 (1999)
68. E Parizet, N Hamzaoui, G Sabatie, Comparison of some listening test methods: a case study. Acta Acustica U Acustica. **91**(2), 356–364 (2005)
69. P Hatziantoniou, J Mourjopoulos, J Worley, in *118th Audio Engineering Society Convention*. Subjective assessments of real-time room dereverberation and loudspeaker equalization, (Barcelona, 2005)
70. Y Hu, PC Loizou, Subjective comparison and evaluation of speech enhancement algorithms. Speech Commun. **49**, 588–601 (2007)
71. K Kondo, *Subjective Quality Measurement of Speech, Its Evaluation, Estimation and Applications*. (Springer, Berlin Heidelberg, 2012)
72. PC Loizou, G Kim, Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions. IEEE Trans. on Audio, Speech, and Lang. Process. **19**(1), 47–56 (2011)
73. H Wang, R Hu, W Tu, C Zhang, The perceptual and statistics characteristic of spatial cues and its application. Int. J. Comput. Sci. Issues. **10**(3), 621–626 (2013)
74. S Bech, N Zacharov (eds.), *Perceptual Audio Evaluation—Theory, Method and Application* (John Wiley & Sons, Chichester, England, 2006), pp. 29–38. Chap. Fundamentals of experimentation
75. ITU-R. *Recommendation BS.1534-1, Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems*. (International Telecommunications Union Radiocommunication Assembly, 2003)
76. ITU-T. *Recommendation P.835, Subjective Test Methodology for Evaluating Speech Communication Systems that Include Noise Suppression Algorithm*. (International Telecommunications Union, Telecommunications Standardization Sector

77. ITU-T. *Recommendation P.800.1, Mean Opinion Score (MOS) Terminology*. International Telecommunications Union, Telecommunications Standardization Sector, 2003)
78. G Halfacree, E Upton, *Raspberry Pi User Guide*, 1st edn. (John Wiley & Sons, Chichester, 2012)
79. Mathworks: MatLab & Simulink: Simulink Reference R2016a. The MathWorks Inc. (2016). The Mathworks Inc. http://www.mathworks.com/
80. M Azarpour, J Siska, G Enzner, in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*. Realtime binaural speech enhancement demo on Raspberry Pi, (New Orleans, 2017)
81. H Levitt, M Bakke, J Kates, A Neuman, T Schwander, M Weiss, Signal processing for hearing impairment. Scand. Audiol. Suppl. **38**, 7–19 (1993)
82. ITU-T Recommendation P.832, Subjective performance evaluation of hands-free terminals (05/2000) (2000)
83. MJ Ball, C Code (eds.), *Instrumental Clinical Phonetics* (Whurr Publishers, London, 1997)
84. NM Razali, YB Wah, Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. J. Stat. Model. Anal. **2**(1), 21–33 (2011)
85. WH Kruskal, WA Wallis, Use of ranks in one-criterion variance analysis. J. Am. Stat. Assoc. **47**(260), 583–621 (1952)