Binaural Speech Enhancement with Spatial Cue Preservation Utilising Simultaneous Masking

Andreas I. Koutrouvelis*, Jesper Jensen†‡, Meng Guo‡, Richard C. Hendriks* and Richard Heusdens*
e-mails: {a.koutrouvelis, r.c.hendriks, r.heusdens}@tudelft.nl and {jesj, megu}@oticon.com
*Circuits and Systems (CAS) Group, Delft University of Technology, the Netherlands
†Dept. Electronic Systems, Aalborg University, Denmark, ‡Oticon A/S, Denmark

***Abstract*—Binaural multi-microphone noise reduction methods aim at noise suppression while preserving the spatial impression of the acoustic scene. Recently, a new binaural speech enhancement method was proposed which chooses per time-frequency (TF) tile either the enhanced target or a suppressed noisy version. The selection between the two is based on the input SNR per TF tile. In this paper we modify this method such that the selection mechanism is based on the output SNR. The proposed modification of deciding which TF tile is target-or noise-dominated leads to choices, which are better aligned with simultaneous masking properties of the auditory system, and, hence, improves the performance over the initial version of the algorithm.**

***Index Terms*—Binaural hearing aids, noise reduction, simultaneous masking.**

I. Introduction

The rapidly increasing communication capabilities between small portable devices make the notion of binaural noise reduction (BNR) [1] increasingly tractable for wireless collaborative hearing aids (HAs) [2]. BNR methods aim at acoustic noise suppression, using the microphones from both HAs, without altering the spatial impression of the acoustic scene.

Typically, BNR methods consist of two beamformers (one at the left and one at the right HA) and, optionally, a post-filter applied to the outputs of the two beamformers for further noise suppression [1]. The BNR methods can be roughly grouped into two main categories: a) methods that require estimates of the relative acoustic transfer functions (RATFs) of all present sources (e.g., [3]–[7]), and b) methods which require *only* the estimated RATF of the target (e.g., [8]–[11]). In this paper we focus on the second category of BNR methods mainly due to the practicality of only relying on the target RATF.

The binaural minimum variance distortionless response (BMVDR) beamformer [5] consists of two MVDR beamformers [12], [13] and requires only an estimate of the noise cross-power spectral density matrix and the RATF of the target. It provides the maximum noise reduction performance within the class of binaural linearly constrained distortionless minimum variance beamformers [5], [6]. However, this is at the cost of distorting the binaural cues of the interferers [5], [6], which will coincide with the binaural cues of the target after processing [5].

The BMVDR-N method, initially proposed in [8] and further investigated in [14], combines the output of the BMVDR with a portion of the noisy unprocessed signal to preserve the binaural cues of the noise. A slightly different approach was presented in [10], referred to as the selective binaural beamformer (SBB). This method uses either the BMVDR output *or* a suppressed version of the unprocessed noisy acoustic scene, depending on whether the target or the noise is dominant in a time-frequency (TF) tile. This classification of target-dominant and noise-dominant TF tiles is accomplished using an estimate of the *input* SNR.

All aforementioned approaches have in common that they intend to preserve the spatial cues of all sources without taking the notion into account that some sources are actually inaudible *after* processing. In this paper we introduce the idea of speech enhancement with binaural cue preservation only of the sources that are audible at the output of the filter. The general advantage of this approach is that degrees of freedom which in traditional approaches are assigned to cue preservation of sources, which turn out to be inaudible after processing (and hence masked) are now released and maybe assigned to noise reduction. More specifically we apply this concept to a modification of the SBB approach. Instead of using the input SNR, we use an estimate of the BMVDR output SNRs at left and right ears [15] for the binary classification. This allows us to better control the characteristics of the noise reaching the ears of the user. Moreover, the proposed method is better aligned with masking properties than the SBB method. If the noise, after processing with the BMVDR beamformer, is inaudible in a TF tile, there is no need to preserve its binaural cues in this specific TF tile and, therefore, the maximum possible noise reduction is achieved by applying the BMVDR. On the other hand, if the noise after processing is audible, the binaural cue distortions introduced by the BMVDR may be audible and, therefore, a scaled version of the noisy acoustic scene is used instead.

II. Notation and Signal Model

We assume for convenience that the two HAs have an equal number of m microphones with $M = 2m$ microphones in total. Without any loss of generality we assume that there is a single target point source and one interferering point source present in the acoustic scene. Stacking all microphone

This work was supported by the Oticon Foundation and the Dutch Technology Foundation STW.

frequency-domain elements into vectors, we have the following signal model for a single TF tile

$$\mathbf{y}(t,f) = \mathbf{x}(t,f) + \underbrace{\mathbf{n}(t,f) + \mathbf{v}(t,f)}_{\mathbf{z}(t,f)} \in \mathbb{C}^{M\times 1}, \qquad (1)$$

where $\mathbf{y}(t,f)$, $\mathbf{x}(t,f)$, $\mathbf{n}(t,f)$, $\mathbf{v}(t,f)$ and $\mathbf{z}(t,f)$ are the noisy, target, interferer, background noise and overall noise vectors for the DFT bin f and time frame t, respectively. The 1-st and the M-th microphones are selected as reference microphones[1] and the corresponding elements of all vectors in Eq. (1) have subscripts L and R, respectively, for notational convenience. Note that $\mathbf{x}(t,f) = \mathbf{a}(t,f)s(t,f)$ and $\mathbf{n}(t,f) = \mathbf{b}(t,f)u(t,f)$, where $\mathbf{a}(t,f)$ and $\mathbf{b}(t,f)$ are the acoustic transfer functions (ATFs) of the target and the interferer, respectively, while $s(t,f)$ and $u(t,f)$ are the target signal and interfering signal at the original positions, respectively.

The BNR methods consists of two filters $\mathbf{w}_L(t,f)$, $\mathbf{w}_R(t,f) \in \mathbb{C}^{M\times 1}$ that are applied to the noisy vector $\mathbf{y}(t,f)$, obtaining the following two outputs

$$\hat{x}_L(t,f) = \mathbf{w}_L^H(t,f)\mathbf{y}(t,f), \quad \hat{x}_R(t,f) = \mathbf{w}_R^H(t,f)\mathbf{y}(t,f),$$

where $\mathbf{w}_L(t,f)$, and $\mathbf{w}_R(t,f)$ are estimated using all microphone recordings from both HAs.

A. Binaural Spatial Information Measures

The binaural spatial information for point sources is measured in terms of the interaural level differences (ILDs) and the interaural phase differences (IPDs). The input/output ILDs and IPDs of the interferer for a single TF tile are given by[2]

$$\text{IPD}^{\text{in}}_{\mathbf{n}} = \angle\frac{b_L}{b_R} \quad \text{and} \quad \text{IPD}^{\text{out}}_{\mathbf{n}} = \angle\frac{\mathbf{w}_L^H\mathbf{b}}{\mathbf{w}_R^H\mathbf{b}}, \qquad (2)$$

$$\text{ILD}^{\text{in}}_{\mathbf{n}} = \left|\frac{b_L}{b_R}\right|^2 \quad \text{and} \quad \text{ILD}^{\text{out}}_{\mathbf{n}} = \left|\frac{\mathbf{w}_L^H\mathbf{b}}{\mathbf{w}_R^H\mathbf{b}}\right|^2. \qquad (3)$$

Similar expressions to Eqs. (2) and (3) exist for the target source. In addition, we quantify binaural spatial characteristics of the background noise in terms of the input and output magnitude square coherence (MSC) [5], [14] given by

$$\text{MSC}^{\text{in}} = \left|\frac{c^{\text{in}}_{LR}}{\sqrt{\left(c^{\text{in}}_{LL}\right)\left(c^{\text{in}}_{RR}\right)}}\right|^2, \ \text{MSC}^{\text{out}} = \left|\frac{c^{\text{out}}_{LR}}{\sqrt{\left(c^{\text{out}}_{LL}\right)\left(c^{\text{out}}_{RR}\right)}}\right|^2, \qquad (4)$$

respectively, where $c^{\text{in}}_{LR} = \mathbf{e}_L^T\mathbf{P}_{\mathbf{v}}\mathbf{e}_R$, $c^{\text{in}}_{LL} = \mathbf{e}_L^T\mathbf{P}_{\mathbf{v}}\mathbf{e}_L$, $c^{\text{in}}_{RR} = \mathbf{e}_R^T\mathbf{P}_{\mathbf{v}}\mathbf{e}_R$, $c^{\text{out}}_{LR} = \mathbf{w}_L^H\mathbf{P}_{\mathbf{v}}\mathbf{w}_R$, $c^{\text{out}}_{LL} = \mathbf{w}_L^H\mathbf{P}_{\mathbf{v}}\mathbf{w}_L$, $c^{\text{out}}_{RR} = \mathbf{w}_R^H\mathbf{P}_{\mathbf{v}}\mathbf{w}_R$, $\mathbf{P}_{\mathbf{v}}$ is the cross-power spectral density matrix of the background noise for a single TF tile, $\mathbf{e}_L^T = [1\ 0, \cdots, 0]$ and $\mathbf{e}_R^T = [0, \cdots, 0\ 1]$. A desired property of a BNR method is to have small MSC, IPD and ILD errors, defined as

$$\text{MSC}^{\text{error}}(t,f) = \left|\text{MSC}^{\text{out}}(t,f) - \text{MSC}^{\text{in}}(t,f)\right|, \qquad (5)$$

$$\text{IPD}^{\text{error}}_{\mathbf{n}}(t,f) = \left|\text{IPD}^{\text{out}}_{\mathbf{n}}(t,f) - \text{IPD}^{\text{in}}_{\mathbf{n}}(t,f)\right|/\pi, \qquad (6)$$

$$\text{ILD}^{\text{error}}_{\mathbf{n}}(t,f) = \left|\text{ILD}^{\text{out}}_{\mathbf{n}}(t,f) - \text{ILD}^{\text{in}}_{\mathbf{n}}(t,f)\right|. \qquad (7)$$

It is only relevant to measure the aforementioned spatial errors of the residual noise in a TF tile, (t,f), when the residual noise is audible at the output. To reflect to which extent the processed noise is masked by the processed target we apply a weighting to the ILD, IPD and MSC errors.

The weights are computed based on the simultaneous masking principle [16] as follows. First the k-th critical band SNR (CBSNR) output with respect to the left and right reference microphones are computed. The left CBSNR is given by

$$\text{CBSNR}_{k,L}(t) = \frac{\sum_{f\in\text{CB}_k}\mathbf{w}_L^H(t,f)\mathbf{P}_{\mathbf{x}}(t,f)\mathbf{w}_L(t,f)}{\sum_{f\in\text{CB}_k}\mathbf{w}_L^H(t,f)\mathbf{P}_{\mathbf{z}}(t,f)\mathbf{w}_L(t,f)}, \qquad (8)$$

where CB_k denotes the index set of DFT bins corresponding to the k-th critical band, and $\mathbf{P}_{\mathbf{x}}(t,f)$ is the cross-power spectral density matrix of the target at the TF tile (t,f). A similar expression exists for the right CBSNR, $\text{CBSNR}_{k,R}(t)$. Then, the weights associated with the k-th critical band are computed. Specifically, the weights for the left reference microphone are given by

$$\phi_{k,L}(t) = \begin{cases} 1, & \text{CBSNR}_{k,L}(t) \le \lambda \\ 1 - \frac{\text{CBSNR}_{k,L}(t)-\lambda}{\rho-\lambda}, & \lambda < \text{CBSNR}_{k,L}(t) < \rho\ , \\ 0, & \text{CBSNR}_{k,L}(t) \ge \rho \end{cases} \qquad (9)$$

where $\lambda = -4$ dB and $\rho = 24$ dB are the noise-tone and tone-noise masking thresholds [16]. If $\text{CBSNR}_{k,L}(t) \ge 24$, the target masks completely the noise at the left reference microphone in the k-th critical band, while if $\text{CBSNR}_{k,L}(t) \le -4$, the noise completely masks the target [16]. The weights at the right reference microphone are computed as in Eq. (9), but using $\text{CBSNR}_{k,R}(t)$ instead of $\text{CBSNR}_{k,L}(t)$.

The average masking-weighted spatial information error measures for the left reference microphone are defined as

$$\text{AvMSC}^{\text{error}}_L = \frac{\sum_{t=1}^T\sum_{k=1}^N\phi_{k,L}(t)\sum_{f\in\text{CB}_k}\text{MSC}^{\text{error}}(t,f)}{\sum_{t=1}^T\sum_{k=1}^N\sum_{f\in\text{CB}_k}\phi_{k,L}(t)},$$

$$\text{AvIPD}^{\text{error}}_L = \frac{\sum_{t=1}^T\sum_{k=1}^N\phi_{k,L}(t)\sum_{f\in\text{CB}_k}\text{IPD}^{\text{error}}_{\mathbf{n}}(t,f)}{\sum_{t=1}^T\sum_{k=1}^N\sum_{f\in\text{CB}_k}\phi_{k,L}(t)},$$

$$\text{AvILD}^{\text{error}}_L = \frac{\sum_{t=1}^T\sum_{k=1}^N\phi_{k,L}(t)\sum_{f\in\text{CB}_k}\text{ILD}^{\text{error}}_{\mathbf{n}}(t,f)}{\sum_{t=1}^T\sum_{k=1}^N\sum_{f\in\text{CB}_k}\phi_{k,L}(t)},$$

with T the number of time-frames and N the number of critical bands. Similar expressions exist for the right reference microphone.

III. Proposed Method

Similarly to the SBB method [10], the proposed method consists of two processing phases: a) the classification phase

[1] The BNR methods aim at preserving the binaural cues of all sources with respect to the reference microphones.

[2] These measures/quantities as well as other measures/quantities introduced in the sequel of the paper are time-frequency varying, however for notational convenience the TF indices (t,f) in some occasions are omitted.

of TF tiles into target-dominant and noise-dominant, and b) the enhancement phase where the BMVDR is applied to the target dominant TF-tiles, while in the noise-dominant TF tiles a scaled (with $0 \leq g \leq 1$) version of the noisy signal is used in both HAs. Let the left and right input narrowband SNRs (NBSNRs) be given by [15]

$$\eta_L^{\text{in}} = \frac{\mathbf{e}_L^T \mathbf{P}_\mathbf{x} \mathbf{e}_L}{\mathbf{e}_L^T \mathbf{P}_\mathbf{z} \mathbf{e}_L}, \quad \eta_R^{\text{in}} = \frac{\mathbf{e}_R^T \mathbf{P}_\mathbf{x} \mathbf{e}_R}{\mathbf{e}_R^T \mathbf{P}_\mathbf{z} \mathbf{e}_R}, \tag{10}$$

respectively. The left and right BMVDR output NBSNRs are given by [15]

$$\eta_L^{\text{out}} = \eta_L^{\text{in}} \left(\mathbf{a}_L^H \mathbf{P}_L^{-1} \mathbf{a}_L \right), \quad \eta_R^{\text{out}} = \eta_R^{\text{in}} \left(\mathbf{a}_R^H \mathbf{P}_R^{-1} \mathbf{a}_R \right), \tag{11}$$

respectively, and $\mathbf{a}_L = (1/a_L)\mathbf{a}$, $\mathbf{a}_R = (1/a_R)\mathbf{a}$, $\mathbf{P}_L^{-1} = P_{\mathbf{z},(1,1)} \mathbf{P}_\mathbf{z}^{-1}$, and $\mathbf{P}_R^{-1} = P_{\mathbf{z},(M,M)} \mathbf{P}_\mathbf{z}^{-1}$, where $P_{\mathbf{z},(1,1)}$ and $P_{\mathbf{z},(M,M)}$ are the first and last diagonal elements, respectively, of $\mathbf{P}_\mathbf{z}$. The filters of the proposed method at the left and right HAs for a single TF tile are given by

$$\mathbf{w}_{\text{Prop.},L} = \begin{cases} \mathbf{w}_{\text{MV},L}, & \eta_L^{\text{out}} \geq \tau, \text{ and } \eta_R^{\text{out}} \geq \tau \\ g\mathbf{e}_L, & \text{otherwise} \end{cases}, \tag{12}$$

$$\mathbf{w}_{\text{Prop.},R} = \begin{cases} \mathbf{w}_{\text{MV},R}, & \eta_L^{\text{out}} \geq \tau, \text{ and } \eta_R^{\text{out}} \geq \tau \\ g\mathbf{e}_R, & \text{otherwise} \end{cases}, \tag{13}$$

with $\mathbf{w}_{\text{MV,L}}$ and $\mathbf{w}_{\text{MV,R}}$ the left and right BMVDR filters, respectively, η_L^{out} and η_R^{out} the output NBSNRs at the left and right reference microphones, respectively, and τ the threshold value which is fixed over frequency and time. The BMVDR filters are given by [5]

$$\mathbf{w}_{\text{MV},L} = \frac{\mathbf{P}_\mathbf{z}^{-1} \mathbf{a} a_L^*}{\mathbf{a}^H \mathbf{P}_\mathbf{z}^{-1} \mathbf{a}}, \quad \mathbf{w}_{\text{MV},R} = \frac{\mathbf{P}_\mathbf{z}^{-1} \mathbf{a} a_R^*}{\mathbf{a}^H \mathbf{P}_\mathbf{z}^{-1} \mathbf{a}}, \tag{14}$$

with $\mathbf{P}_\mathbf{z}$ the cross-power spectral density matrix of the total noise, and a_L and a_R the two reference elements of $\mathbf{a}$.

A. Improvements of the SBB method

In our evaluation, we compare our proposed method to an improved version of the SBB method. The improvements consider two aspects. First, unlike the original SBB [10] which uses only one input NBSNR in the classification stage, our implementation of SBB uses both η_L^{in} and η_R^{in} in order to guarantee target dominance in both ears. Secondly, in the original SBB method [10], the scaling parameter g was selected as

$$g = \min \left(\frac{1}{\mathbf{w}_{\text{MV,L}}^H \mathbf{P}_\mathbf{z} \mathbf{w}_{\text{MV,L}}}, \frac{1}{\mathbf{w}_{\text{MV,R}}^H \mathbf{P}_\mathbf{z} \mathbf{w}_{\text{MV,R}}} \right). \tag{15}$$

Computing g with Eq. (15) might, in some situations, boost the noise. Instead, in this paper we select g as

$$g = \min \left(\sqrt{\frac{\mathbf{w}_{\text{MV,L}}^H \mathbf{P}_\mathbf{z} \mathbf{w}_{\text{MV,L}}}{\mathbf{e}_L^T \mathbf{P}_\mathbf{z} \mathbf{e}_L}}, \sqrt{\frac{\mathbf{w}_{\text{MV,R}}^H \mathbf{P}_\mathbf{z} \mathbf{w}_{\text{MV,R}}}{\mathbf{e}_R^T \mathbf{P}_\mathbf{z} \mathbf{e}_R}} \right), \tag{16}$$

in both the proposed and the SBB methods.

As in [10] we use an average g (computed across the noise-dominated DFT bins) for each time-frame for both the proposed and the SBB methods to mitigate coloration of the residual noise. Hence, g is time-varying but constant over frequency.

B. Basic Principle

There are two main reasons to use η_L^{out} and η_R^{out} (the proposed method) instead of η_L^{in} and η_R^{in} (the SBB method), in the classification stage. First, the main goal of the proposed method is to achieve the maximum possible noise suppression, without altering the binaural cues of the *audible* processed noise. Therefore, if the processed noise is masked by the processed target, there is no reason to preserve any binaural cues of the noise and, then, the largest possible noise reduction is achieved by using the BMVDR output.

Secondly, judging whether the noise is masked by the target is easier if this is done *after* processing (based on η_L^{out} and η_R^{out}) than *before* processing (based on η_L^{in} and η_R^{in}). This is because, after processing, the binaural cues of the noise coincides with the binaural cues of the target and one can use the monaural simultaneous masking principle described in [16]. Moreover, after processing, masking becomes independent of the spatial layout of the sources in the acoustic scene.

Based on the aforementioned two facts, the proposed method will be more robust than the SBB method to changing acoustical scenarios assuming that a fixed threshold τ is used in both methods. This will be shown in Sections III-C, III-D.

C. Example 1: Point Noise Source

Fig. 1 demonstrates the difference between the proposed method and the SBB method, for a synthetic speech shaped target source in the front (0 degrees), an interfering speech shaped noise source to the right (-80 degrees) and a small amount of microphone self noise. Figs. 1(a) and 1(b) depict the estimated input and output NBSNRs at the left and right reference microphones, respectively. Figs. 1(c) and 1(d) show the $\text{AvIPD}_L^{\text{error}}$ and $\text{AvIPD}_R^{\text{error}}$ of the interferer vs. the output segmental SNR (SSNR) for the two methods, respectively, over a threshold value, τ, ranging from -50 dB to 50 dB with a step-size of 0.5 dB. Figs. 1(e) and 1(f) show the $\text{AvILD}_L^{\text{error}}$ and $\text{AvILD}_R^{\text{error}}$ of the interferer vs the output SSNR, respectively, for the same range of τ values. The output SSNR at the left reference microphone is defined as

$$\text{SSNR}_L^{\text{out}} = \frac{1}{T} \sum_{t=1}^{T} 10\log_{10} \frac{||\mathbf{q}_{t,L}||_2^2}{||\hat{\mathbf{q}}_{t,L} - \mathbf{q}_{t,L}||_2^2}, \tag{17}$$

with $\mathbf{q}_{t,L}$ the time-frame t of the clean target signal at the left reference microphone, $\hat{\mathbf{q}}_{t,L}$ its estimate. A similar expression holds for the $\text{SSNR}_R^{\text{out}}$.

Let us examine four interesting τ values for this specific example. If $\tau > 29$ dB, both SBB and the proposed method will not achieve any noise suppression, but they will simply scale the noisy signal by g. This is because, $\eta_L^{\text{in}}, \eta_R^{\text{in}}, \eta_L^{\text{out}}, \eta_R^{\text{out}} <$ 29 dB for all frequency bins. Thus, the values of the performance curves in Figs. 1(c,d,e,f) corresponding to $\tau > 29$ dB will be in the left bottom corner.

If $\tau = 22.5$ dB, most parts of the $\eta_L^{\text{out}}, \eta_R^{\text{out}}$ curves will be above $\tau = 22.5$ dB, while all the frequency bins of the curves $\eta_L^{\text{in}}, \eta_R^{\text{in}}$ will be below $\tau = 22.5$ dB. This means that the proposed method will achieve some noise reduction, while the SBB method will not suppress the noise at all. Moreover, since $\tau = 22.5$, the processed noise in all the frequency bins that correspond to $\eta_L^{\text{out}} > 22.5, \eta_R^{\text{out}} > 22.5$ will be almost inaudible and, therefore, the weighted average binaural cue errors will be approximately zero. In conclusion, a) none of the methods caused any audible binaural cue errors, b) the proposed method achieved some noise reduction, while the SBB method did not achieve any noise reduction. In Figs. 1(c,d,e,f), the performances for $\tau = 22.5$ dB are shown with a red □ marker and a blue ○ marker for the proposed method and the SBB method, respectively.

For $\tau = 2$ dB there will be some frequency bins (in the region 7-8 kHz) of η_L^{in}, and η_R^{in} that will be above $\tau = 2$ dB as well. The number of these frequency bins will be much less than the number of the frequency bins of $\eta_L^{\text{out}}, \eta_R^{\text{out}}$ that will be above $\tau = 2$ dB. Thus, the proposed method will achieve larger amount of noise reduction. Both methods will cause audible binaural cue errors for $\tau = 2$.

For values $\tau < -8$ dB both methods will have identical performance, i.e., both methods will apply the BMVDR beamformer to all frequency bins. This corresponds to the top right corner (marked with a black star), in Figs. 1(c,d,e,f).

It is clear that the proposed method achieves a better output SSNR than the SBB for many values of $\text{AvILD}_L^{\text{error}}$, $\text{AvILD}_R^{\text{error}}$, $\text{AvIPD}_L^{\text{error}}$ and $\text{AvIPD}_R^{\text{error}}$ errors, in this acoustic scenario.

D. Example 2: Diffuse Noise

Similarly to Fig. 1, Fig. 2 shows the difference between the proposed method and the SBB method when there is a target speech shaped source at the front (0 degrees), a diffuse noise field and a small amount of microphone self noise. As mentioned in Section II-A, a proper measure for binaural spatial distortions in diffuse noise fields is the $\text{AvMSC}_L^{\text{error}}$ and $\text{AvMSC}_R^{\text{error}}$ errors. Therefore, in Fig. 2, we use the $\text{AvMSC}_L^{\text{error}}$ $\text{AvMSC}_R^{\text{error}}$ errors to show the performance difference between the two methods.

It is worth noting that in Figs. 2(a,b) the curves η_L^{out} and η_L^{in} have very similar structure, i.e., they are approximately vertically shifted. The same applies also for the curves η_R^{out} and η_R^{in}. This means the two methods will give more or less identical SSNR for any AvMSC error. This can be observed in Figs. 2(c,d), were the performance curves are very similar.

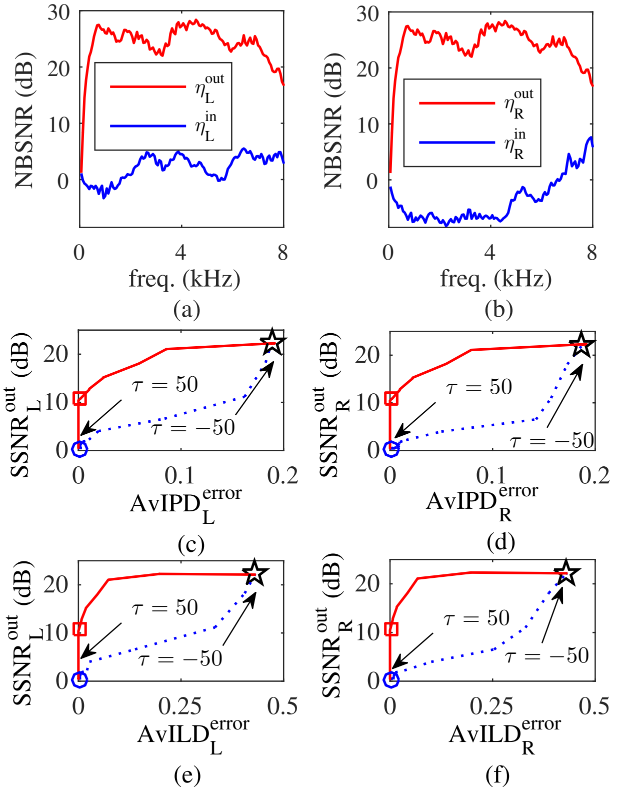

Fig. 1. Simulation example 1 comparing the proposed (red) with the BSS (blue) method and the BMVDR (black star). For $\tau = 22.5$, the performance of the proposed and the BSS method is illustrated with a red □ marker and a blue ○ marker, respectively.

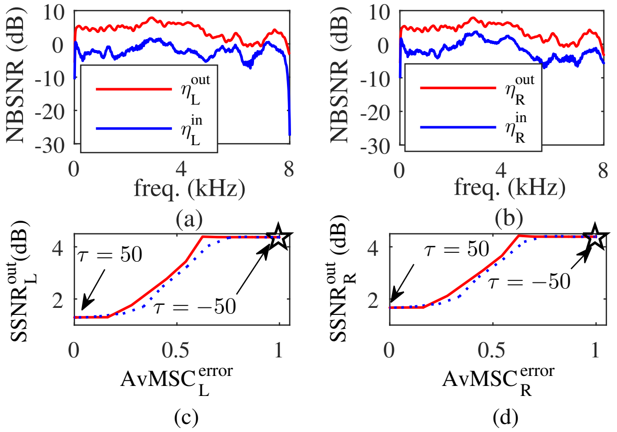

Fig. 2. Simulation example 2 comparing the proposed (red) with the BSS (blue) method and the BMVDR (black star).

IV. Simulations

In this section, the proposed method is compared with the SBB method [10] for $\tau = -50 : 0.5 : 50$ dB, and the BMVDR-N method [8], [14] with $N = 0 : 0.1 : 1$. The comparison is done in two different noisy acoustic scenarios. In the first scenario the noise component is a single interferer (a male talker) on the right of the HA user (at -80 degrees). In the second scenario the noise component is diffuse noise which is created using different speech shape noise realizations from 72 different angles around the head. In both scenarios, the target is a female talker positioned in the front (i.e., 0 degrees) of the HA user, and the microphone self-noise (in all microphones) is 50 dB smaller with respect to the target signal at the left reference microphone. In both simulated scenarios, we used the anechoic head impulse responses from [17] to simulate both the point sources and the diffuse noise. The female and male talker point sources were placed 0.8 m from

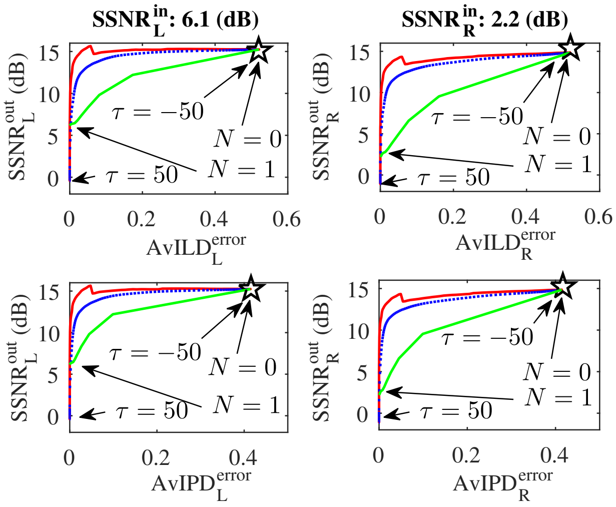

Fig. 3. Scenario 1 comparing the proposed (red) with the BSS (blue) method, the BMVDR-N (green) and the BMVDR (black star).

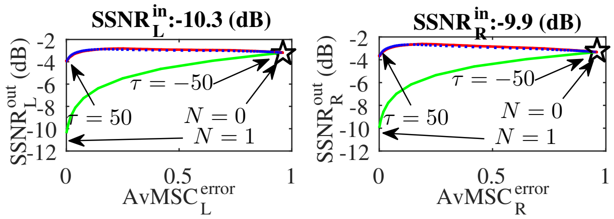

Fig. 4. Scenario 2 comparing the proposed (red) with the BSS (blue) method, the BMVDR-N (green) and the BMVDR (black star).

the head, while the point sources that are for the diffuse noise are placed 3 m from the center of the head. All simulated signals have a duration of 14 seconds in which the first 4 seconds the noise is active only. The BMVDR filters used the true $\mathbf{a}$ and an estimate of $\mathbf{P_z}$ using a perfect VAD. The $\eta_L^{\text{in}}, \eta_R^{\text{in}}, \eta_L^{\text{out}}$, and η_R^{out} are estimated using the method in [15] using a perfect VAD and the true $\mathbf{a}$. We used an overlap and add methodology for processing the signals with a frame size of 10 ms and overlap 50%. The sampling frequency is 16 kHz.

Figs. 3 and 4 show a performance comparison for the first and second simulated acoustic scenario, respectively. The gap in performance, between the SBB method and the proposed method, depends on the input/output NBSNR structure and type of the noise field as discussed in Sections III-C and III-D. For the first simulated acoustical scenario, the proposed method achieves a higher noise reduction performance (as measured with SSNR) for most binaural spatial error values. This is due to the big difference of the structure of the output NBSNR compared to the structure of the input NBSNR as explained in Section III-C. However, this is not the case for the second simulated scenario as expected (see Section III-D), since the structure of the output NBSNR is very similar with the structure of the input NBSNR. Moreover, note that the BMVDR-N method has the worst performance over the other two methods in all acoustic scenarios for most N values.

V. Conclusion

We proposed a modified version of the selective binaural beamformer (SBB) approach. The proposed method differs from the SBB approach in the classification stage of the time-frequency (TF) tiles. It uses the output SNR for labeling the TF tiles either to target-dominant or noise-dominant. This modification is better aligned with the simultaneous masking principle. Furthermore, it was experimentally shown that in some acoustical scenarios the proposed method provides larger amount of noise reduction than BSS for the same binaural spatial distortions.

References

[1] S. Doclo, W. Kellermann, S. Makino, and S. Nordholm, “Multichannel signal enhancement algorithms for assisted listening devices,” *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 18–30, Mar. 2015.

[2] J. M. Kates, *Digital hearing aids*. Plural publishing, 2008.

[3] B. Cornelis, S. Doclo, T. Van den Bogaert, M. Moonen, and J. Wouters, “Theoretical analysis of binaural multimicrophone noise reduction techniques,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 2, pp. 342–355, Feb. 2010.

[4] E. Hadad, S. Doclo, and S. Gannot, “The binaural LCMV beamformer and its performance analysis,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 24, no. 3, pp. 543–558, Mar. 2016.

[5] E. Hadad, D. Marquardt, S. Doclo, and S. Gannot, “Theoretical analysis of binaural transfer function MVDR beamformers with interference cue preservation constraints,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 23, no. 12, pp. 2449–2464, Dec. 2015.

[6] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, “Relaxed binaural LCMV beamforming,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 25, no. 1, pp. 137–152, Jan. 2017.

[7] A. I. Koutrouvelis, R. C. Hendriks, J. Jensen, and R. Heusdens, “Improved multi-microphone noise reduction preserving binaural cues,” in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Mar. 2016.

[8] T. Klasen, T. Van den Bogaert, M. Moonen, and J. Wouters, “Binaural noise reduction algorithms for hearing aids that preserve interaural time delay cues,” *IEEE Trans. Signal Process.*, vol. 55, no. 4, pp. 1579–1585, Apr. 2007.

[9] J. Thiemann, M. Müller, and S. van de Par, “A binaural hearing aid speech enhancement method maintaining spatial awareness for the user,” in *EURASIP Europ. Signal Process. Conf. (EUSIPCO)*, Sep. 2014, pp. 321–325.

[10] J. Thiemann, M. Müller, D. Marquardt, S. Doclo, and S. van der Par, “Speech enhancement for multimicrophone binaural hearing aids aiming to preserve the spatial auditory scene,” *EURASIP J. Advances Signal Process.*, 2016.

[11] H. As’ad, M. Bouchard, and H. Kamkar-Parsi, “Perceptually motivated binaural beamforming with cues preservation for hearing aids,” in *IEEE Canadian Conf. Electrical and Computer Engineering (CCECE)*, May 2016.

[12] J. Capon, “High-resolution frequency-wavenumber spectrum analysis,” *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.

[13] B. D. Van Veen and K. M. Buckley, “Beamforming: A versatile approach to spatial filtering,” *IEEE ASSP Mag.*, vol. 5, no. 5, pp. 4–24, Apr. 1988.

[14] D. Marquardt, “Development and evaluation of psychoacoustically motivated binaural noise reduction and cue preservation techniques,” Ph.D. dissertation, Carl von Ossietzky Universität Oldenburg, 2015.

[15] J. Jensen and M. S. Pedersen, “Analysis of beamformer directed single-channel noise reduction system for hearing aid applications,” in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Apr. 2015, pp. 5728–5732.

[16] T. Painter and A. Spanias, “Perceptual coding of digital audio,” *Proceedings of the IEEE*, vol. 88, no. 4, pp. 451–515, Apr. 2000.

[17] H. Kayser, S. Ewert, J. Annemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, “Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses,” *EURASIP J. Advances Signal Process.*, vol. 2009, pp. 1–10, Dec. 2009.