Binaural Speaker Localization Integrated Into an Adaptive Beamformer for Hearing Aids

Mehdi Zohourian, *Student Member, IEEE*, Gerald Enzner, *Senior Member, IEEE*, and Rainer Martin, *Fellow, IEEE*

***Abstract*—In this paper, we present and compare novel algorithms to localize simultaneous speakers using four microphones distributed on a pair of binaural hearing aids. The framework consists of two groups of localization algorithms, namely, beamforming-based and statistical model based localization algorithms. We first generalize our previously proposed methods based on beamforming techniques to the binaural configuration with 2×2 microphones. Next, we contribute two statistical model based methods for binaural localization using the maximum likelihood approach that also takes head-related transfer functions and unknown noise conditions into account. The methods enable the localization of multiple source positions for all azimuth angles and do not require prior training of binaural cues. The proposed localization algorithms are integrated into a generalized side-lobe canceller (GSC) to extract the desired speaker in the presence of competing speakers and background noise and when the head of the listener turns. The GSC components are adapted with the frequency-wise target presence probability and the frame-wise broadband direction-of-arrival (DOA) estimates that track the turns of the listener's head. We evaluate the performance of the localization algorithms individually and also in the context of the adaptive binaural beamformer in various noisy and reverberant conditions. Finally, we introduce a new adaptive beamformer, which combines the GSC with multichannel speech presence probability estimation and achieves superior source separation performance in noisy environment.**

***Index Terms*—Binaural source localization, beamforming, source separation, hearing-aids.**

I. Introduction

BINAURAL speech enhancement algorithms are used in various head-mounted communication devices, e.g., hearing aids (HAs). Modern HAs are equipped with multiple microphones and the left and right devices are connected through a wireless link. The spatial diversity provided by these multiple microphones enables the attenuation of the directional interfering signals such as competing speakers and are superior to single-channel approaches. Suppression of interfering signals can be grouped into two families of algorithms: 1- Blind source separation (BSS) algorithms that exploit the statistical characteristics of all involved sources. For instance, independent component analysis (ICA) based algorithms use second or higher order statistics of the signals [1], [2]. 2-Beamforming-based algorithms that rely on a known microphone configuration and a spatial model of the signal. In fact, beamformers are spatio-temporal filters that are steered towards the desired source and extract it in the presence of competing speakers and ambient noise [3], [4].

For beamforming-based approaches prior knowledge of the DOA of both the desired and interfering signals are crucial to achieve a reliable performance for the separation of concurrent speakers [4]. Some contributions assume a perfectly known DOA of the target direction whereas the others estimate DOAs using binaural localization algorithms. The latter can also account for a mismatch of the steering vector due to the movement of the head of both speakers and listener. For instance, adaptive differential beamformers [5] are typically applied in HAs. They allow a good suppression of interferences in the back half plane with the assumption that the target signal is in front of the listener. This scheme has been further improved by steering the look direction of the beamformer to other angles than 0° [6] or integrated with spatial noise reduction for the suppression of directional noise [7]. Furthermore, the linearly-constrained minimum-variance (LCMV) or as a special case the minimum-variance distortionless response (MVDR) beamformers [8] and also the multichannel Wiener filter (MWF) [9], [10] have been employed successfully in HAs. Recently, some of these binaural beamforming approaches address the problem of binaural cue preservation [11], [12]. The well-known *generalized side-lobe canceller* (GSC) [13] is an adaptive implementation of the LCMV beamformer [14] and is widely used in practice. It consists of three main components, a fixed beamformer that steers toward the desired source, a blocking matrix that provides noise and interference references, and an adaptive noise canceler that cancels the residual noise in the output of the fixed beamformer. This structure has been further improved in more sophisticated scenarios to prevent signal degradation [15] and to deal with non stationary noise [16]. To improve the estimation of the blocking matrix, [4] has introduced a model-based GSC which is integrated with a localization algorithm using a free-field linear array of microphones. Localization algorithms

Manuscript received August 10, 2017; revised November 21, 2017; accepted November 30, 2017. Date of publication December 11, 2017; date of current version January 8, 2018. This work was supported in part by the People Programme (Marie Curie Actions) of the European Unions Seventh Framework Programme FP7/2007-2013/ under REA Grant PITN-GA-2012-31752 (ICanHear), and in part by the European Fund for Regional Development under Grant EFRE-0800372 (RaVis-3D). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Hsin-min Wang. *(Corresponding author: Mehdi Zohourian.)*

The authors are with the Institute of Communication Acoustics, Ruhr-Universität Bochum, Bochum 44780, Germany (e-mail: mehdi.zohourian@rub.de; gerald.enzner@rub.de; rainer.martin@rub.de).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2017.2782491

for free-field arrays often optimize a cost function that considers only the relative phase of the microphone signals. In binaural localization, however, the two major binaural cues, namely, *interaural time/phase difference* (ITD/IPD) and the *interaural level difference* (ILD) are taken into account. The ILD plays an important role in binaural DOA estimation especially at high frequencies [17].

The impact of using joint IPD and ILD cues in binaural localization have been broadly studied. Some state-of-the-art methods are based on prior training of the binaural cues corresponding to individual source position through a statistical model [18], [19]. Other approaches derive cost functions based on the comparison between the estimated and the prototype *head-related transfer functions* (HRTFs) [17]. The prototype could be extracted from either a head model or a database. For instance, the dual delay line approach and its variants [20], [21] equalize the binaural signals with HRTFs from a database for each possible source position. In [21] the estimated relative transfer function (RTF) are compared with the prototype RTFs from the database. In these approaches the DOA estimation of only a single source were integrated with the MVDR beamformer to extract the desired source in the presence of background noise. In [22], however, the binaural cues are modeled by a Gaussian mixture model which is used to create a soft mask for the separation of two speakers at each time-frequency bin. In [23] blindly estimated impulse responses were aligned with an IPD/ILD model to achieve the binaural localization task. The authors in [24] use the well-known MUSIC algorithm to localize multiple sources by exploitation of low-rank frequency bins as detected by the coherence test. Furthermore, some approaches resolve *front-back confusion* using the information from head movements which are tracked using an inertial measurement sensor [25].

In this work we aim at the localization and separation of multiple speakers under varying background noise. We use *behind-the-ear* (BTE) HAs with 2×2 microphones. We propose two groups of binaural localization algorithms, namely, beamforming-based and statistical model-based approaches. These approaches have been studied broadly for the free-field array. However, we apply them to the binaural configuration taking the head shadowing effect into account. In the beamforming-based approaches the beamformer steers the target-beam or the null-beam to all possible source position and select the candidates maximizing or minimizing the response power. On the other hand, the model-based approaches assume a statistical representation of the underlying signals and use maximum likelihood (ML) optimization. Although the ML method is computationally complex, it can potentially work well for arbitrary propagation conditions. In [26] the ML approach has also been employed to estimate the DOA, however, in an informed way given the noise-free signal by the third wirelessly-linked microphone. In this work, however, we utilize the ML approach in an unsupervised manner.

We integrate the localization algorithm in an adaptive binaural beamformer using the GSC structure devised in [4]. We adapt the system to the binaural configuration taking the ILD cues into account. The binaural cues are characterized in the form of HRTFs and are extracted from a database [27]. We also track the listener's head turns using online broadband DOA estimation. Without any prior training our algorithm is able to localize and separate concurrent speakers that are located in the horizontal plane around the listener. Finally, based on the ML localization approach we develop a new adaptive beamformer that integrates the GSC with a multichannel speech presence probability (SPP) algorithm. The SPP improves the source separation performance by modeling the noise as an additional position-dependent component in the estimation of target presence probability.

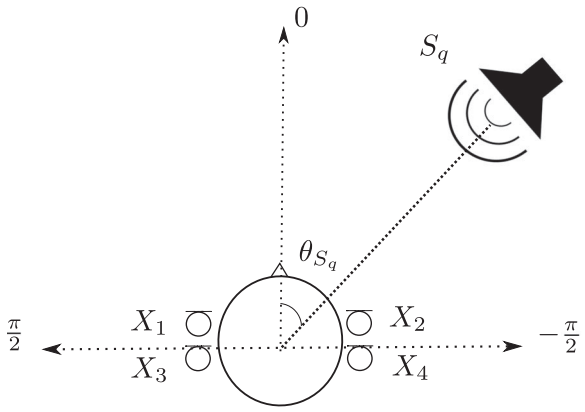


Fig. 1. Binaural configuration, a source S_q, and the coordinate system.

The remainder of the paper is organized as follows. Section II describes the multi-channel signal model used in this paper. Sections III and IV discuss the beamforming-based and model-based binaural localization algorithms, respectively. In Section V we present the adaptive binaural beamformer for source separation. Section VI introduces a multichannel SPP estimation algorithm which is combined with the adaptive beamformer. Experimental evaluations are illustrated in Section VII. Section VIII concludes this paper.

II. Binaural Signal Model

In the scenario depicted in Fig. 1 we consider binaural signals from Q sources received by $M = 4$ microphones distributed on a pair of binaural HAs. Using the convolution operator $*$ the received signal at each microphone $m \in \{1, ..., M\}$ is written as

$$x_m(n) = \sum_{q=1}^{Q} s_q(n) * h_{qm}(n) + \nu_m(n), \tag{1}$$

where $s_q(n)$ represents the q-th point source signal, $h_{qm}(n)$ indicates a binaural room impulse response (BRIR) from source q to microphone m, $\nu_m(n)$ is the noise at microphone m, and n is the sampling index. To analyze signals in the STFT domain, we take a $K-$point discrete Fourier transform (DFT) on overlapped and windowed signal frames. Using matrix notation we thus obtain

$$\boldsymbol{X}(k,b) = \boldsymbol{H}(k,\boldsymbol{\Theta})\boldsymbol{S}(k,b) + \boldsymbol{V}(k,b), \tag{2}$$

where (k, b) indicate frequency and frame indices. In principle, $\boldsymbol{H}$ is the matrix of binaural room transfer functions (BRTFs) of

the M microphones determined by

$$\boldsymbol{H}(k,\boldsymbol{\Theta}) = [\boldsymbol{H}_1(k,\theta_1), \boldsymbol{H}_2(k,\theta_2), ..., \boldsymbol{H}_Q(k,\theta_Q)], \quad (3)$$

where $\boldsymbol{H}_q(k,\theta_q) = [H_{q1}(k,\theta_q), H_{q2}(k,\theta_q), .., H_{qM}(k,\theta_q)]^T$. Here, θ_q is the azimuth location of source q. The signal vectors are given by

$$\begin{aligned}\mathbf{X}(k,b) &= [X_1(k,b), X_2(k,b), .., X_M(k,b)]^T \\ \mathbf{S}(k,b) &= [S_1(k,b), S_2(k,b), .., S_Q(k,b)]^T \\ \mathbf{V}(k,b) &= [V_1(k,b), V_2(k,b), .., V_M(k,b)]^T. \end{aligned} \quad (4)$$

In this work, we aim at the estimation of $\boldsymbol{\Theta} = [\theta_1, \theta_2, ..., \theta_Q]^T$ and the corresponding clean signals $\boldsymbol{S}$. However, we do not manage to estimate room impulse responses perfectly. Therefore, some of the spectral characteristics of room still persists in the output of the beamformer.

It is worth noting that due to the sparsity and disjointness property of speech signals, it is commonly assumed that only one source is predominantly active at each time-frequency bin. Therefore, for all proposed approaches we search for a single source at each time-frequency bin and discard possible combinations of Q sources in the optimization process. In other words, we consider $\boldsymbol{H}(k,\theta_q)$ instead of $\boldsymbol{H}(k,\boldsymbol{\Theta})$.

III. Beamforming-Based Localization Algorithm

Beamforming-based localization approaches are attractive due to their simplicity. They search for possible source locations that either maximize the power of the *target-beam* or minimize the power of the *null-beam*. We term the former as the *target beamforming* (TBF) and the latter as the *null-steering beamformer* (NBF) techniques. In the current work we extend our previously proposed beamforming-based localization algorithms [28] to 2×2 HA microphones. This increases the spatial diversity and is helpful to localize sources at all azimuth angles.

A. Target Beamforming (TBF)

A well-known method in this category is *steered-response power* the (SRP) algorithm [29]. Generally, SRP is a filter-and-sum beamformer (FSB) that scans all azimuth angles and searches for the candidates maximizing its output power. Although its optimization relies on broadband information, it is formulated in the frequency domain. When the output of the beamformer $\boldsymbol{W}(k,\theta)$ is denoted by $\hat{S} = \boldsymbol{W}^H \boldsymbol{X}$, the output power of the beamformer is given by

$$\Phi_{\hat{s}\hat{s}}(k) = \boldsymbol{W}^H(k,\theta)\boldsymbol{\Phi}_{\boldsymbol{XX}}(k)\boldsymbol{W}(k,\theta). \quad (5)$$

Here, $\boldsymbol{\Phi}_{\boldsymbol{XX}}(k) = E\left\{\boldsymbol{X}(k)\boldsymbol{X}^H(k)\right\}$ is the spatial covariance matrix which is estimated by employing a first-order recursive system for smoothing over successive frames as

$$\hat{\boldsymbol{\Phi}}_{\boldsymbol{XX}}(k,b) = \alpha\hat{\boldsymbol{\Phi}}_{\boldsymbol{XX}}(k,b-1) + (1-\alpha)\boldsymbol{X}(k,b)\boldsymbol{X}^H(k,b), \quad (6)$$

with a smoothing parameter $0 < \alpha < 1$.

The original SRP approach is based on a delay-and-sum beamformer which does not model the head shadowing effect. Therefore, we incorporate joint IPD and ILD into a FSB to improve the localization performance. Since the HRTFs have a non-uniform power for different azimuth angles which affects the steered-response power non-uniformly, and the effect of ILD should be preserved in each channel, we propose to use the matched filter approach [30] with a unit norm as

$$\boldsymbol{W}_{TBF}(k,\theta) = \frac{\boldsymbol{H}(k,\theta)}{||\boldsymbol{H}(k,\theta)||}. \quad (7)$$

Similar to the spectral weighting of the SRP method indicated in [31], the cost function is weighted by $||\boldsymbol{X}(k)||^2$ which balances the weights of low and high frequency contributions in DOA estimates. This improves the robustness of the broadband DOA estimation against reverberation and deemphasizes undesired local maxima. Therefore, the TBF cost function is determined by

$$\Lambda_{TBF}(k,\theta) = \frac{\boldsymbol{H}^H(k,\theta)\hat{\boldsymbol{\Phi}}_{\boldsymbol{XX}}(k)\boldsymbol{H}(k,\theta)}{||\boldsymbol{H}(k,\theta)||^2||\boldsymbol{X}(k)||^2}. \quad (8)$$

The cost function is maximized across all position candidates θ. Note that this method implicitly assumes spatially uncorrelated noise signals at each pair of microphones.

B. Null-Steering Beamforming (NBF)

The next approach is based on a null-steering beamformer where the null-steering vector first equalizes the binaural signals. Then, the method scans the room and searches for the candidates minimizing its output power. We utilize this idea in the binaural localization context which was also done earlier in [21], [32] and extend it in this work to account for all azimuth angles. The cost function is determined by

$$\Lambda_{NBF}(k,\theta) = \sum_{m,m'} \bar{\boldsymbol{W}}^H_{NBF}(k,\theta)\boldsymbol{\Phi}^{(m,m')}_{\boldsymbol{XX}}(k)\bar{\boldsymbol{W}}_{NBF}(k,\theta), \quad (9)$$

where $\boldsymbol{\Phi}^{(m,m')}_{\boldsymbol{XX}}(k)$ is the spatial covariance matrix of a pair of microphones $\{m,m'\}$ and is estimated by temporal smoothing. $\bar{\boldsymbol{W}}_{NBF}(k,\theta) = [W_m(k,\theta), W_{m'}(k,\theta)]^T$ is the null-steering beamformer considering only two sensors.

Since the magnitude of HRTFs is involved in the localization, it is of great importance to equalize the signals. Therefore, the NBF filters are designed considering the cross-relation technique where each pair of microphones is filtered with contralateral HRTFs for each possible source position. Moreover, we apply the constraint $||\bar{\boldsymbol{W}}_{NBF}(k,\theta)|| = 1$ to normalize the response power of the beamformer. This leads to a more robust broadband DOA estimates. Therefore, we obtain the energy-normalized NBF filter vector as

$$\begin{aligned}\bar{\boldsymbol{W}}_{NBF}(k,\theta) = &\frac{1}{\sqrt{|H_m(k,\theta)|^2 + |H_{m'}(k,\theta)|^2}} \\ &\times [H_{m'}(k,\theta), -H_m(k,\theta)]^H. \end{aligned} \quad (10)$$

This cost function is also normalized by the magnitudes $|X_m(k)||X_{m'}(k)|$ of the microphone signals which are independent of estimated binaural cues and affect the estimated power

similar to the (PHAT) weighting in the SRP-PHAT method [29]. It increases the robustness of the broadband DOA estimation by reducing the sensitivity to environmental conditions such as reverberation. For the sake of consistency we add a negative sign to the cost function and maximize it across all angles θ,

$$\Lambda_{NBF}(k,\theta) = -\sum_{m,m'} \frac{\bar{\boldsymbol{W}}_{NBF}^{H}(k,\theta)\boldsymbol{\Phi}_{\boldsymbol{XX}}^{(m,m')}(k)\bar{\boldsymbol{W}}_{NBF}(k,\theta)}{|X_m(k)||X_{m'}(k)|}. \tag{11}$$

IV. Model-Based Localization Algorithms

The presented beamforming-based localization algorithms do not consider noise characteristics and seems to be suboptimal when the noise signal is spatially coherent. To cope with this problem we need to develop an approach which takes the full signal model into account. The ML DOA estimator is an important class of this type of algorithms. It requires the probability density function of the signals under consideration. This technique has been extensively used for the problem of DOA estimation in free-field scenarios [33], [34]. In this work we employ it for binaural localization. Two classes of ML approaches have been investigated depending on the model assumption which are described below.

A. Deterministic Maximum Likelihood (DML) DOA Estimation

If we assume that the source signal is unknown and deterministic and that the noise samples follow a zero-mean complex Gaussian distribution, the probability distribution of a narrowband signal $\boldsymbol{X}$ for each frequency is given by

$$P(\boldsymbol{X}|\theta, S, \boldsymbol{\Phi}_{\boldsymbol{VV}}) = \frac{1}{\pi^M |\boldsymbol{\Phi}_{\boldsymbol{VV}}|} \exp\left(-(\boldsymbol{X}-\boldsymbol{H}S)^H \boldsymbol{\Phi}_{\boldsymbol{VV}}^{-1}(\boldsymbol{X}-\boldsymbol{H}S)\right). \tag{12}$$

In this equation $\boldsymbol{\Phi}_{\boldsymbol{VV}}$ is the covariance matrix of the noise DFT coefficients and $|.|$ denotes the determinant operation. Note that for readability the frequency index k and the DOA index θ is dropped in (12) and in the equations below.

Under the assumption that a sequence of DFT frames of microphone signals, i.e., $\boldsymbol{X}^B = [\boldsymbol{X}(1), ..., \boldsymbol{X}(b), ..., \boldsymbol{X}(B)]^T$ and the sequence of DFT frames of the narrowband clean signals, i.e., $\boldsymbol{S}^B = [S(k), ..., S(b), ..., S(B)]^T$ are temporarily independent and identically distributed (i.i.d.), we obtain

$$P(\boldsymbol{X}^B|\theta, \boldsymbol{S}^B, \boldsymbol{\Phi}_{\boldsymbol{VV}}) = \prod_{b=1}^{B} P(\boldsymbol{X}(b)|\theta, S(b), \boldsymbol{\Phi}_{\boldsymbol{VV}}). \tag{13}$$

Therefore, the log-likelihood function for (13) is

$$L(\boldsymbol{X}^B|\theta, \boldsymbol{S}^B, \boldsymbol{\Phi}_{\boldsymbol{VV}}) = -BM\log\pi - B\log|\boldsymbol{\Phi}_{\boldsymbol{VV}}| - \sum_{b=1}^{B}(\boldsymbol{X}(b)-\boldsymbol{H}S(b))^H \boldsymbol{\Phi}_{\boldsymbol{VV}}^{-1}(\boldsymbol{X}(b)-\boldsymbol{H}S(b)). \tag{14}$$

In principle, we factor the noise covariance matrix as $\boldsymbol{\Phi}_{\boldsymbol{VV}} = \sigma^2\boldsymbol{\Gamma}_{\boldsymbol{VV}}$, where σ^2 is called the power of the noise signal that captures its spectral characteristics. Moreover, $\boldsymbol{\Gamma}_{\boldsymbol{VV}}$ is called the normalized spatial covariance matrix of the noise signal that captures the spatial characteristics of the noise signal. Hence, the log-likelihood function (14) is expanded to

$$L(\boldsymbol{X}^B|\theta, \boldsymbol{S}^B, \boldsymbol{\Phi}_{\boldsymbol{VV}}) = -BM\log\pi - B\log|\boldsymbol{\Gamma}_{\boldsymbol{VV}}| - BM\log\sigma^2 - \frac{1}{\sigma^2}\sum_{b=1}^{B}(\boldsymbol{X}(b)-\boldsymbol{H}S(b))^H \boldsymbol{\Gamma}_{\boldsymbol{VV}}^{-1}(\boldsymbol{X}(b)-\boldsymbol{H}S(b)). \tag{15}$$

The DML approach first estimates the power of the noise signal $\hat{\sigma}^2$ and estimates the clean signal $\hat{S}$ conditioned on the DOA parameter θ. Then, it substitutes the estimated parameters in (15) and maximizes it w.r.t. θ. The DOA is thus estimated by maximizing

$$\Lambda_{DML}(k,\theta) = \frac{1}{B}\sum_{b=1}^{B} \boldsymbol{X}^H(k,b)\boldsymbol{\Gamma}_{\boldsymbol{VV}}^{-1}(k)\boldsymbol{P}_{\boldsymbol{H}}(k,\theta)\boldsymbol{X}(k,b), \tag{16}$$

where

$$\boldsymbol{P}_{\boldsymbol{H}}(k,\theta) = \boldsymbol{H}(k,\theta)\left(\boldsymbol{H}^H(k,\theta)\boldsymbol{\Gamma}_{\boldsymbol{VV}}^{-1}(k)\boldsymbol{H}(k,\theta)\right)^{-1} \times \boldsymbol{H}^H(k,\theta)\boldsymbol{\Gamma}_{\boldsymbol{VV}}^{-1}(k), \tag{17}$$

has the properties of an orthogonal projection matrix. The proof of (16) is presented in more detail in Appendix.

If we rearrange the terms in (16) and use non-recursive temporal averaging for the estimation of the covariance matrix,

$$\hat{\boldsymbol{\Phi}}_{\boldsymbol{XX}}(k) = \frac{1}{B}\sum_{b=1}^{B}\boldsymbol{X}(k,b)\boldsymbol{X}^H(k,b), \tag{18}$$

and normalize the sample covariance matrix to its norm we can write the DOA cost function as

$$\Lambda_{DML}(k,\theta) = \frac{\boldsymbol{H}^H(k,\theta)\boldsymbol{\Gamma}_{\boldsymbol{VV}}^{-1}(k)\hat{\boldsymbol{\Phi}}_{\boldsymbol{XX}}(k)\boldsymbol{\Gamma}_{\boldsymbol{VV}}^{-1}(k)\boldsymbol{H}(k,\theta)}{\boldsymbol{H}^H(k,\theta)\boldsymbol{\Gamma}_{\boldsymbol{VV}}^{-1}(k)\boldsymbol{H}(k,\theta)\boldsymbol{X}^H(k)\boldsymbol{X}(k)}. \tag{19}$$

Under the assumption that the noise signal is spatially uncorrelated, i.e., $\boldsymbol{\Gamma}_{\boldsymbol{VV}} = \boldsymbol{I}_{M\times M}$, (17) simplifies to $\boldsymbol{P}_{\boldsymbol{H}} = \boldsymbol{H}(\boldsymbol{H}^H\boldsymbol{H})^{-1}\boldsymbol{H}^H$ which is known as the projection matrix to the signal subspace. The cost function is simplified to [31]

$$\Lambda_{DML}^{Uncorr}(k,\theta) = \frac{\boldsymbol{H}^H(k,\theta)\hat{\boldsymbol{\Phi}}_{\boldsymbol{XX}}(k)\boldsymbol{H}(k,\theta)}{||\boldsymbol{H}(k,\theta)||^2||\boldsymbol{X}(k)||^2}, \tag{20}$$

which is identical to the solution of the target beamforming method presented in (8). In fact, the DML algorithm can be interpreted as the extension of the TBF approach considering the ambient noise characteristics.

B. Stochastic Maximum Likelihood (SML) DOA Estimation

If we assume that the source signal is a stochastic random process, the noise is a stationary process, and both follow Gaussian distributions, we may write the probability density function

TABLE I
A SUMMARY OF COST FUNCTIONS OF VERSATILE LOCALIZATION ALGORITHMS

algorithms	cost functions $\Lambda(k,\theta) =$	Equation
TBF	$\frac{\boldsymbol{H}^H(k,\theta)\hat{\boldsymbol{\Phi}}_{\boldsymbol{XX}}(k)\boldsymbol{H}(k,\theta)}{\|\boldsymbol{H}(k,\theta)\|^2\|\boldsymbol{X}(k)\|^2}$	(8)
NBF	$-\sum_{m,m'}\frac{\bar{\boldsymbol{W}}_{NBF}^H(k,\theta)\boldsymbol{\Phi}_{\boldsymbol{XX}}^{(m,m')}(k)\bar{\boldsymbol{W}}_{NBF}(k,\theta)}{\lvert X_m(k)\rvert\lvert X_{m'}(k)\rvert}$	(11)
DML	$\frac{\boldsymbol{H}^H(k,\theta)\boldsymbol{\Gamma}_{\boldsymbol{VV}}^{-1}(k)\hat{\boldsymbol{\Phi}}_{\boldsymbol{XX}}(k)\boldsymbol{\Gamma}_{\boldsymbol{VV}}^{-1}(k)\boldsymbol{H}(k,\theta)}{\boldsymbol{H}^H(k,\theta)\boldsymbol{\Gamma}_{\boldsymbol{VV}}^{-1}(k)\boldsymbol{H}(k,\theta)\boldsymbol{X}^H(k)\boldsymbol{X}(k)}$	(19)
SML	$-\log\lvert\boldsymbol{P}_{\boldsymbol{H}}(\theta)\hat{\boldsymbol{\Phi}}_{\boldsymbol{XX}}(k)\boldsymbol{P}_{\boldsymbol{H}}^H(\theta)+\hat{\sigma}^2\boldsymbol{P}_{\boldsymbol{H}}^{\perp}(\theta)\boldsymbol{\Gamma}_{\boldsymbol{VV}}(k)\rvert$	(26)

of the narrowband signal $\boldsymbol{X}(k)$ as

$$P(\boldsymbol{X}|\theta,\boldsymbol{\Phi}_{\boldsymbol{XX}}) = \frac{1}{\pi^M|\boldsymbol{\Phi}_{\boldsymbol{XX}}|}\exp\left(-\boldsymbol{X}^H\boldsymbol{\Phi}_{\boldsymbol{XX}}^{-1}\boldsymbol{X}\right). \quad (21)$$

Similar to the DML approach we assume that a sequence of DFT frames of the narrowband signal is temporarily independent and identically distributed (i.i.d.). We thus write the log-likelihood function as

$$\begin{aligned} L(\boldsymbol{X}^B|\theta,\boldsymbol{\Phi}_{\boldsymbol{XX}}) &= \log\prod_{b=1}^{B}P(\boldsymbol{X}(b)|\theta,\boldsymbol{\Phi}_{\boldsymbol{XX}}) \\ &= -BM\log\pi - B\log|\boldsymbol{\Phi}_{\boldsymbol{XX}}| - B\,\mathrm{Tr}\left\{\boldsymbol{\Phi}_{\boldsymbol{XX}}^{-1}\hat{\boldsymbol{\Phi}}_{\boldsymbol{XX}}\right\}, \end{aligned} \quad (22)$$

where $\hat{\Phi}_{\boldsymbol{XX}}$ is the estimation of the spatial covariance matrix using non-recursive averaging given by (18). According to the signal model in (2) and based on the disjointness property of speech DFT coefficients, the spatial covariance matrix of the microphone signals is written as

$$\boldsymbol{\Phi}_{\boldsymbol{XX}}(k) = \boldsymbol{H}(k,\theta)\boldsymbol{H}^H(k,\theta)\Phi_{SS}(k) + \sigma^2\boldsymbol{\Gamma}_{\boldsymbol{VV}}(k), \quad (23)$$

where Φ_{SS} and σ^2 denote the power of the clean and noise signals, respectively. The SML approach first derives the estimation of the power of the clean and noise signals $\hat{\Phi}_{SS}$ and $\hat{\sigma}^2$ conditioned on the DOA parameter θ. Then, it substitutes the estimated parameters in (22) and maximizes it w.r.t. θ. The power of the clean signal is estimated by [34]

$$\hat{\Phi}_{SS} = \frac{\boldsymbol{H}^H(k,\theta)\boldsymbol{\Gamma}_{\boldsymbol{VV}}{}^{-1}(k)(\hat{\boldsymbol{\Phi}}_{\boldsymbol{XX}}(k)-\hat{\sigma}^2\boldsymbol{\Gamma}_{\boldsymbol{VV}}(k))\boldsymbol{\Gamma}_{\boldsymbol{VV}}^{-1}(k)\boldsymbol{H}(k,\theta)}{\left(\boldsymbol{H}^H(k,\theta)\boldsymbol{\Gamma}_{\boldsymbol{VV}}^{-1}(k)\boldsymbol{H}(k,\theta)\right)^2}. \quad (24)$$

The estimation of the power of the noise is obtained as [34]

$$\hat{\sigma}^2 = \frac{1}{M-Q}\mathrm{Tr}\left(\boldsymbol{P}_{\boldsymbol{H}}^{\perp}(k,\theta)\hat{\boldsymbol{\Phi}}_{\boldsymbol{XX}}(k)\boldsymbol{\Gamma}_{\boldsymbol{VV}}^{-1}(k)\right). \quad (25)$$

In this equation $\boldsymbol{P}_{\boldsymbol{H}}^{\perp}(k,\theta) = \boldsymbol{I}_{M\times M} - \boldsymbol{P}_{\boldsymbol{H}}(k,\theta)$, where $\boldsymbol{P}_{\boldsymbol{H}}(k,\theta)$ is given by (17).

Therefore, by substituting (24), (25) in (23) and inserting (23) in (22) we arrive at the DOA cost function [34]

$$\begin{aligned} &\Lambda_{SML}(k,\theta) = \\ &\quad -\log|\boldsymbol{P}_{\boldsymbol{H}}(\theta)\hat{\boldsymbol{\Phi}}_{\boldsymbol{XX}}(k)\boldsymbol{P}_{\boldsymbol{H}}^H(\theta) + \hat{\sigma}^2\boldsymbol{P}_{\boldsymbol{H}}^{\perp}(\theta)\boldsymbol{\Gamma}_{\boldsymbol{VV}}(k)|, \end{aligned} \quad (26)$$

which is maximized across all position candidates.

Table I summarizes the DOA cost functions for all proposed methods. For all methods the narrowband DOA estimation is achieved by maximizing the cost functions at each frequency bin across all possible source locations, denoted as

$$\hat{\theta}(k,b) = \underset{\theta}{\mathrm{argmax}}\left(\Lambda_x(k,\theta)\right). \quad (27)$$

The broadband DOA estimation is determined by summing the cost functions across all frequencies and then taking the global maximum and discard any other local maxima, i.e.,

$$\hat{\theta}(b) = \underset{\theta}{\mathrm{argmax}}\sum_{k=1}^{K}\Lambda_x(k,\theta). \quad (28)$$

This improves the robustness of the source estimates in noisy and reverberant conditions.

Note that for all introduced localization cost functions we replace the true binaural room transfer functions $\boldsymbol{H}(k,\theta)$ with the head-related transfer function prototypes $\hat{\boldsymbol{H}}(k,\theta)$ that are extracted from a database [27]. Moreover, for the case of an ideal spherically isotropic (diffuse) noise field each element of the matrix $\boldsymbol{\Gamma}(k)$ is estimated by averaging P HRTFs across the all azimuth angles in the horizontal plane [35],

$$\hat{\Gamma}_{mn}(k) = \sum_{p=1}^{P}\hat{H}_m(k,\theta_p)\hat{H}_n^*(k,\theta_p). \quad (29)$$

V. ADAPTIVE BINAURAL BEAMFORMING

In our experiments the binaural localization algorithm is integrated in an adaptive beamformer using the GSC structure [13]. Fig. 2 shows the block diagram of the adaptive binaural beamformer used to extract the desired source q. The GSC consists of a fixed beamformer $\mathbf{W}_{f_q}(k,b)$, an adaptive blocking matrix $\mathbf{B}_q(k,b)$, and an adaptive noise canceller $\mathbf{W}_{V_q}(k,b)$. Previously, this structure had been integrated with the SRP algorithm [29] to separate concurrent speakers using a linear array of microphones [4]. We adapted this beamformer for HA

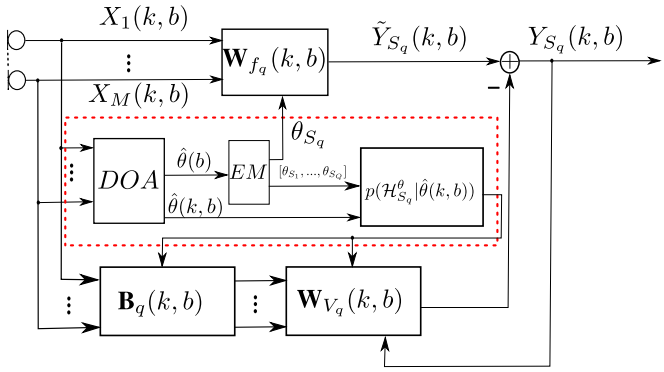


Fig. 2. Block diagram of the proposed adaptive binaural beamformer used for the extraction of the desired source q.

microphones to separate speakers in the frontal hemisphere [32]. We achieved similar performance to [4] despite the reduction in the number of microphones and the larger relative distance between the microphones. We showed in [32] that the lower spatial diversity in the binaural configuration as compared to the linear array of five microphones [4] can be compensated to a large extent if the head-shadowing effect is properly taken into account. In the current work we adapt the beamformer to 2×2 BTE microphones and test it to separate simultaneous speakers that could be located in all azimuth angles. Each part of the beamformer is described below.

1) Target Presence Probability (TPP): The narrowband DOA estimations $\hat{\theta}(k,b)$ are summarized in the TPP estimation which is commonly modeled using a Gaussian mixture model (GMM) [4], [36]. Given the DOAs of Q concurrent speakers we may formulate Q hypotheses $\mathcal{H}^{\theta}_{S_q}, q \in \{1,..,Q\}$. Hypothesis $\mathcal{H}^{\theta}_{S_q}(k,b)$ represents the activity of speech source q in the time-frequency bin (k,b). Therefore, the conditional probability of the q-th hypothesis given the estimated DOA may be written as [37]:

$$p(\mathcal{H}^{\theta}_{S_q}(k,b)|\hat{\theta}(k,b)) = \frac{\rho_{S_q}\,\mathcal{N}\left(\hat{\theta}(k,b)|\mu_{S_q},\sigma^2_{S_q}\right)}{\sum_{i=1}^{Q}\rho_{S_i}\,\mathcal{N}\left(\hat{\theta}(k,b)|\mu_{S_i},\sigma^2_{S_i}\right)}, \tag{30}$$

where the mean of each Gaussian component indicates the location of each source, i.e., $\theta_{S_q} = \mu_{S_q}$. In this equation, $\sigma^2_{S_q}$ and ρ_{S_q} denote the variance and the prior probability of the q-th Gaussian component. A well-know method to estimate the GMM parameters is the *expectation-maximization* (EM) algorithm [38]. The authors in [36] implement the EM algorithm in a batch processing format assuming a Dirichlet distribution for the prior probabilities. Their method allows for the estimation of the number of active sources. Another alternative is to use the EM algorithm in a frame-based processing as, for instance, in [4], [39]. In this case the EM algorithm has been implemented over a collection of successive frames to ensure that the position of all active sources are estimated properly.

In our approach we estimate the broadband DOAs $\hat{\theta}(b)$ across the first second of data and cluster them using the EM algorithm. We thus find the initial positions θ_{S_q} of all involved speakers as the mean values of the GMM in (30). We approximate the variances and the priors of the GMM intuitively as $\pi/18$ and $1/Q$, respectively. Then, we compute the posteriori probability of the q-th hypothesis by inserting the estimated narrowband DOA $\hat{\theta}(k,b)$ in (30).

Note that in our implementation the number of active sources is a known parameter in the computation of the TPP. This is critical information especially for the task of source separation in which the extraction of all involved speakers is desired. However, in case only one target speaker is desired and all other speakers are considered as interfering sources, the quality of the extracted target source does not change dramatically if the number of interfering sources increases. This is due to the functionality of the GSC that blocks the target signal and cancels out the non-target signals from the output of the beamformer and is thus independent of the number of interfering sources.

2) Fixed Beamformer: The fixed beamformer $\boldsymbol{W}_{f_q}(k,\theta)$ is designed using the MVDR approach. Therefore, the output of the beamformer is given by

$$\tilde{Y}_{S_q}(k,b) = \frac{\boldsymbol{H}_q^H(k,\theta_q)\boldsymbol{\Gamma}^{-1}_{\boldsymbol{VV}}(k)}{\boldsymbol{H}_q^H(k,\theta_q)\boldsymbol{\Gamma}^{-1}_{\boldsymbol{VV}}(k)\boldsymbol{H}_q(k,\theta_q)}\boldsymbol{X}(k,b). \tag{31}$$

It is shown in (50) that the MVDR beamformer is intrinsic to the DML approach for the estimation of the clean signal. The advantage of this method as compared to the DSB beamformer is that the ILD cues also contribute to the steering vector and thus provide more diversity compared to the DSB.

3) Adaptive Blocking Matrix: The adaptive blocking matrix blocks the target signal and provides a noise and interference reference for the adaptive noise canceller. The idea is to first construct the projection matrix into the target signal subspace. It is estimated using the TPP values multiplied by the normalized covariance matrix of the noisy signal which is also temporarily smoothed to prevent large variations [4]. The projection matrix of the q-th source $\hat{\boldsymbol{P}}_q(k,b)$ is estimated by

$$\begin{aligned}\hat{\boldsymbol{P}}_q(k,b) = &\left(1 - p(\mathcal{H}^{\theta}_{S_q}|\hat{\theta}(k,b)\right)\hat{\boldsymbol{P}}_q(k,b-1)\\ &+ p(\mathcal{H}^{\theta}_{S_q}|\hat{\theta}(k,b))\frac{\hat{\boldsymbol{\Phi}}_{\boldsymbol{XX}}(k,b)}{\|\mathbf{X}(k,b)\|^2}.\end{aligned} \tag{32}$$

Then, the projection to the complementary subspace is computed by

$$\hat{\boldsymbol{P}}_q^{\perp}(k,b) = \boldsymbol{I}_{M\times M} - \hat{\boldsymbol{P}}_q(k,b), \tag{33}$$

where $\boldsymbol{I}_{M\times M}$ is the identity matrix. Since the target signal should be canceled via the blocking matrix, its rank should be reduced by one. Therefore, we compute the blocking matrix of the q-the source $\boldsymbol{B}_q$ using an operator $\kappa_{(M-1)M}(\cdot)$ which selects the first $(M-1)$ rows and M columns of the matrix argument as

$$\boldsymbol{B}_q(k,b) = \kappa_{(M-1)M}\left(\hat{\boldsymbol{P}}_q^{\perp}(k,b)\right). \tag{34}$$

4) Adaptive Noise Canceller: The adaptive noise canceller uses a normalized least mean-square (NLMS) algorithm [4]

$$\boldsymbol{W}_{V_q}(k,b+1) = \boldsymbol{W}_{V_q}(k,b) + \alpha_q \frac{Y^*_{S_q}(k,b)\boldsymbol{B}_q(k,b)\boldsymbol{X}(k,b)}{||\boldsymbol{B}_q(k,b)\boldsymbol{X}(k,b)||^2}, \tag{35}$$

with an adaptive step-size $\alpha_q = \left(1 - p(\mathcal{H}^{\theta}_{S_q}|\hat{\theta}(k,b))\right)\alpha_f$, where α_f denotes a fixed stepsize factor. We can tune α_f to balance the separation performance and the reduction of the estimation error variance.

Then, we have the estimated signal of the q-th source as

$$Y_{S_q} = \left(\boldsymbol{W}^H_{f_q}(k,b) - \boldsymbol{W}^H_{V_q}(k,b)\boldsymbol{B}_q(k,b)\right)\boldsymbol{X}(k,b). \tag{36}$$

VI. Adaptive Binaural Beamformer With Multichannel Speech Presence Probability

In the previous section we developed the DOA-based GSC for binaural source separation in conjunction with four localization algorithms. Among these localization algorithms the SML method provides additional information of the power of the clean speech and noise signals that could be employed for the improvement of the adaptive beamformer.

In this section we propose a new adaptive beamformer that integrates the previously proposed GSC with a multichannel SPP. The SPP improves the estimation of the TPP model described in (30). The previous TPP estimation model assumes only Q clusters modeling the location of Q speakers and fixed priors and variances for the GMM. Now, we allow for an extra component in the GMM (30) that represents the noise with a mean at $\mu_N = \pi$ and a variance of $\sigma^2_N = \pi$. As before, we set the variance of the speaker positions to $\sigma^2_{S_q} = \pi/18$. Then, the priors ρ_{S_q} of the GMM model are estimated using SPP as outlined below.

The SPP estimation has been investigated previously both for the case of single-channel [40] and multichannel [41] speech enhancement. Authors in [41] decompose the covariance of the noise signal into coherent and incoherent parts and derive the generalized likelihood ratio for the target signal only. In the current study we also derive a closed-form equation for multichannel SPPs for the case of multiple sources that can be interpreted as an extension of the single-channel SPP equation. We express the generalized likelihood ratio for both target and interferer signals separately and with respect to the ambient noise. Furthermore, we take the HRTFs into consideration and define the *a posteriori* SNR as the ratio of the output power of the beamformer and the power of the ambient noise. We use the SML method described in Section IV-B to estimate the power of speech and noise signals.

Given Q concurrent speakers in the presence of ambient noise we may statistically formulate $Q+1$ hypotheses

- $\mathcal{H}^{\zeta}_{S_q}$, speech source $q \in \{1,..Q\}$ is present, i.e., $\boldsymbol{X}(k,b) = \boldsymbol{H}_q(k,\theta_q)S_q(k,b) + \boldsymbol{V}(k,b)$
- $\mathcal{H}_V$, all speech sources are absent, i.e., $\boldsymbol{X}(k,b) = \boldsymbol{V}(k,b)$.

Here, $\mathcal{H}^{\zeta}_{S_q}$ is an SNR-based hypothesis and is different to the DOA-based hypothesis $\mathcal{H}^{\theta}_{S_q}$ described before. Moreover, since we consider the disjointness property of speech signals in the spectral domain, the hypotheses of the presence of any mixture of speech sources will not be considered. We thus write the probability of each hypothesis given the noisy signal as

$$p(\mathcal{H}^{\zeta}_{S_q}|X) = \frac{p(X|\mathcal{H}^{\zeta}_{S_q})p(\mathcal{H}^{\zeta}_{S_q})}{\sum_{i=1}^{Q} p(X|\mathcal{H}^{\zeta}_{S_i})p(\mathcal{H}^{\zeta}_{S_i}) + p(X|\mathcal{H}_V)p(\mathcal{H}_V)}, \tag{37}$$

in which $p(\mathcal{H}^{\zeta}_{S_q})$ and $p(\mathcal{H}_V)$ denote prior probabilities of the q-th speech source presence and speech absence, respectively. We may simplify (37) to

$$p(\mathcal{H}^{\zeta}_{S_q}|X) = \frac{\Lambda_q}{1+\sum_{i=1}^{Q}\Lambda_i}, \tag{38}$$

where $\Lambda_q = \frac{p(\mathcal{H}^{\zeta}_{S_q})}{p(\mathcal{H}_V)}\frac{p(X|\mathcal{H}^{\zeta}_{S_q})}{p(X|\mathcal{H}_V)}$ is the generalized likelihood ratio (GLR) for source q.

We assume that DFT coefficients of speech and noise signals follow complex Gaussian distributions and are mutually independent. Similar to (23) we write the spatial covariance matrix of the microphone signals for the q-th hypothesis as $\boldsymbol{\Phi}_{X_qX_q} = \boldsymbol{H}_q\boldsymbol{H}^H_q\Phi_{S_qS_q} + \sigma^2\boldsymbol{\Gamma}_{\boldsymbol{VV}}$. Therefore, the GLR is expressed as

$$\Lambda_q = \frac{p(\mathcal{H}^{\zeta}_{S_q})}{p(\mathcal{H}_V)}\frac{|\sigma^2\boldsymbol{\Gamma}_{\boldsymbol{VV}}|}{|\boldsymbol{\Phi}_{X_qX_q}|}\exp\left(-\boldsymbol{X}^H\left(\boldsymbol{\Phi}^{-1}_{X_qX_q} - \frac{1}{\sigma^2}\boldsymbol{\Gamma}^{-1}_{\boldsymbol{VV}}\right)\boldsymbol{X}\right). \tag{39}$$

In this equation we write

$$\frac{|\sigma^2\boldsymbol{\Gamma}_{\boldsymbol{VV}}|}{|\boldsymbol{\Phi}_{X_qX_q}|} = \frac{\sigma^{2M}|\boldsymbol{\Gamma}_{\boldsymbol{VV}}|}{\sigma^{2M}|\boldsymbol{\Gamma}_{\boldsymbol{VV}}||\boldsymbol{\Gamma}^{-1}_{\boldsymbol{VV}}\boldsymbol{H}_q\boldsymbol{H}^H_q\frac{\Phi_{S_qS_q}}{\sigma^2} + \boldsymbol{I}|} \tag{40}$$

$$= \frac{1}{1+\zeta_q\boldsymbol{H}^H_q\boldsymbol{\Gamma}^{-1}_{\boldsymbol{VV}}\boldsymbol{H}_q}, \tag{41}$$

where $\zeta_q = \frac{\Phi_{S_qS_q}}{\sigma^2}$ is the *a priori* SNR of source q. Using the matrix inversion lemma, we obtain

$$\boldsymbol{\Phi}^{-1}_{X_qX_q} - \frac{1}{\sigma^2}\boldsymbol{\Gamma}^{-1}_{\boldsymbol{VV}} = -\frac{\zeta_q}{\sigma^2}\frac{\boldsymbol{\Gamma}^{-1}_{\boldsymbol{VV}}\boldsymbol{H}_q\boldsymbol{H}^H_q\boldsymbol{\Gamma}^{-1}_{\boldsymbol{VV}}}{1+\zeta_q\boldsymbol{H}^H_q\boldsymbol{\Gamma}^{-1}_{\boldsymbol{VV}}\boldsymbol{H}_q}. \tag{42}$$

Therefore, we can write (39) as

$$\Lambda_q = \frac{p(\mathcal{H}^{\zeta}_{S_q})}{p(\mathcal{H}_V)}\frac{1}{1+\zeta_q\boldsymbol{H}^H_q\boldsymbol{\Gamma}^{-1}_{\boldsymbol{VV}}\boldsymbol{H}_q} \times \exp\left(\frac{\frac{\zeta_q}{\sigma^2}}{1+\zeta_q\boldsymbol{H}^H_q\boldsymbol{\Gamma}^{-1}_{\boldsymbol{VV}}\boldsymbol{H}_q}\boldsymbol{H}^H_q\boldsymbol{\Gamma}^{-1}_{\boldsymbol{VV}}\boldsymbol{X}\boldsymbol{X}^H\boldsymbol{\Gamma}^{-1}_{\boldsymbol{VV}}\boldsymbol{H}_q\right). \tag{43}$$

We define $\delta_q = \boldsymbol{H}^H_q\boldsymbol{\Gamma}^{-1}_{\boldsymbol{VV}}\boldsymbol{H}_q$ which is a scalar and expand the argument inside the exp function by δ^2_q. We thus obtain a term that represents the instantaneous output power of the MVDR

beamformer for source q as

$$\Phi_{\hat{S}_q\hat{S}_q} = \frac{\boldsymbol{H}_q^H \boldsymbol{\Gamma}_{\boldsymbol{VV}}^{-1} \boldsymbol{X}\boldsymbol{X}^H \boldsymbol{\Gamma}_{\boldsymbol{VV}}^{-1} \boldsymbol{H}_q}{\left(\boldsymbol{H}_q^H \boldsymbol{\Gamma}_{\boldsymbol{VV}}^{-1} \boldsymbol{H}_q\right)^2}. \tag{44}$$

Here, we define $\gamma_q = \frac{\Phi_{\hat{S}_q\hat{S}_q}}{\sigma^2}$ as the *a posteriori* SNR of source q. Therefore, the GLR is simplified to

$$\Lambda_q = \frac{p(\mathcal{H}_{S_q}^{\zeta})}{p(\mathcal{H}_V)} \frac{1}{1+\delta_q\zeta_q} \exp\left(\frac{\delta_q^2\zeta_q}{1+\delta_q\zeta_q}\gamma_q\right). \tag{45}$$

For the special case of uncorrelated white noise using a free-field microphone array we have $\delta_q = M$ where M is the number of microphones. Hence, we obtain

$$\Lambda_{q_{uncorr}} = \frac{p(\mathcal{H}_{S_q}^{\zeta})}{p(\mathcal{H}_V)} \frac{1}{1+M\zeta_q} \exp\left(\frac{M^2\zeta_q}{1+M\zeta_q}\gamma_q\right). \tag{46}$$

If $M = 1$ we achieve the well-known expression for the GLR for single-channel noise reduction [8]. The SPPs (37) estimate the priors of the GMM in (30). We use the GSC output power to estimate the *a posteriori* SNR.

VII. Evaluation Results

In the previous sections, we have presented four localization methods and integrated them in the binaural GSC to separate concurrent speakers. We also integrated the GSC with the multichannel SPP which is estimated using the SML algorithm. In this section we evaluate the performance of the localization algorithms individually and also in the context of the adaptive binaural beamformer for the separation of concurrent speakers using instrumental measures. Also, the accuracy of the estimated SPP and its utility for the adaptive beamformer will be investigated.

We perform our experiments in a reverberant room of dimensions 7.5 m × 6.3 m × 3.3 m, a reverberation time of $T_{60} = 0.4$ s and a critical distance of 1.1 m. We use the front and back microphones of a pair of BTE HAs attached to a dummy head. The distance between monaural microphones is 9 mm. Loudspeakers playing male and female utterances were placed at a distance of 1.2 m from the dummy head thus outside the critical distance.

We use speech signals from the TSP database [42]. Each signal has a duration of 80 s consisting of four male and four female speakers. Audio is recorded at 48 kHz and later downsampled to 16 kHz. Signals are segmented using a *Hann* window of length 32 ms with an overlap of 16 ms between successive DFT frames. The number of DFT bins equals 1024. Note that we use HRTF prototypes $\hat{\boldsymbol{H}}(k,\theta)$ from the database [27] which are not matched to the BRIRs of our recordings. The database includes HRIRs for 6 channels from which 4 HRIRs corresponding to front and back microphones of the left and right HAs were selected.

We consider four different background noise types with varying SNRs, namely, no noise, spatially uncorrelated white noise, spatially diffuse white noise, and spatially diffuse babble noise. The diffuse noise signals are recorded in our lab.

We also evaluate the performance of the localization algorithms in conjunction with the adaptive beamformer using the perceptual evaluation of speech quality (PESQ) [43], the short-time objective intelligibility (STOI) [44], and the signal-to-interference ratio (SIR) [45]. For the computation of these measurements the clean speech signal is used as the reference signal.

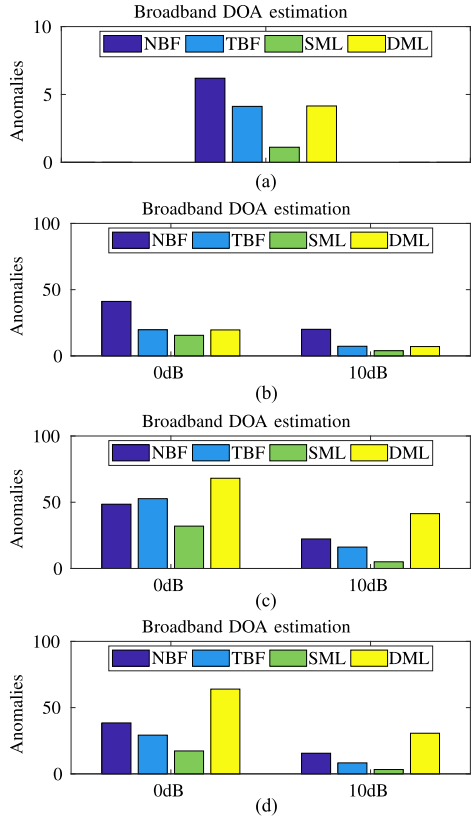


Fig. 3. The broadband DOA estimation of two speakers in a reverberant room (a) without noise, (b) with spatially uncorrelated white noise, (c) with spatially diffuse white noise (d) with spatially diffuse babble noise.

A. Evaluation for a Fixed Head Position

In this experiments the DOA estimates are evaluated in a static manner when the head of the listener is fixed. Experiments are conducted for the two talkers case. One source is always assumed to be at 30°. The other speaker is located at positions on the full azimuth circle starting from 0° and increasing clockwisely with steps of 30°. This results in 11 mixtures. Results for the broadband DOA estimates are reported in terms of the number of outliers (anomalies). The anomalies measure the percentage of frames with DOA estimations outside a preset error interval. The interval for the correct estimation is $\pm 10^\circ$. The limit of 10° is chosen since the spatial resolution of the beamformer does not justify a smaller threshold. Results averaged across all possible source locations are presented in Fig. 3. It is observed that the SML approach achieves the most accurate broadband DOA estimations on average especially in the presence of diffuse noise. As can be seen in Fig. 3, in the presence of diffuse white noise the accuracy of all localization algorithms degrades. The reason lies in the low SNR of high frequency bins that distorts the ILD cues. In addition, the diffuse white noise

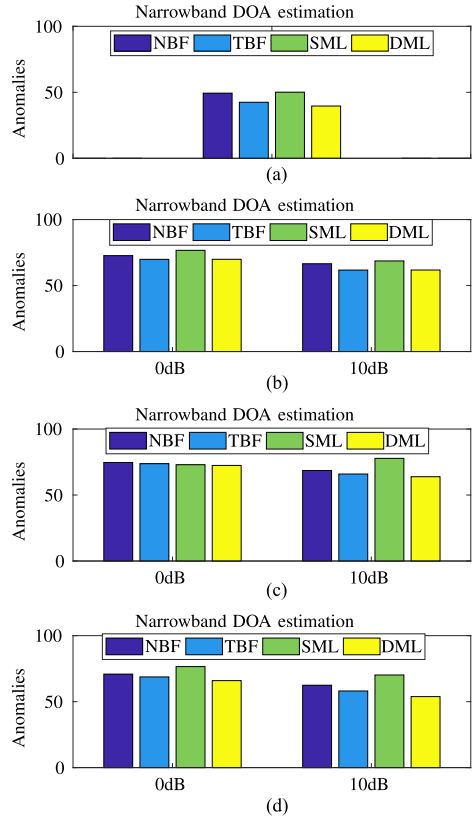


Fig. 4. The narrowband DOA estimation of two speakers in a reverberant room (a) without noise, (b) with spatially uncorrelated white noise, (c) with spatially diffuse white noise (d) with spatially diffuse babble noise.

has a higher coherence in low frequency bins which makes the DOA estimates less accurate than the uncorrelated white noise.

Fig. 4 illustrates the narrowband localization results in terms of anomalies. Here, anomalies measure the percentage of time-frequency bins with absolute DOA estimation errors of more than 10°. Results are averaged across all introduced source positions. As can be seen, the TBF approach achieves better performance than other approaches whereas the SML approach is less robust as compared to other algorithms. The low accuracy of SML is due to its sensitivity to estimation errors in the covariance of the noise as well as the mismatch of HRTFs in low SNR bins. Both lead to an inaccurate regularization factor in the cost function (26) and thus to the lower performance of the SML localization algorithm.

However, the reason why the broadband SML approach outperforms others is that in this technique the effect of noisy bins will be remarkably compensated by reliable ones resulting in a better DOA accuracy. To further analyse this, we compute the broadband DOA estimation using the SML approach as

$$\hat{\theta}_{SML}(b) = \underset{\theta}{\operatorname{argmax}} \sum_{k=1}^{K} -\log|\boldsymbol{\Phi}_{\boldsymbol{XX}}(k)|$$

$$= \underset{\theta}{\operatorname{argmin}} \prod_{k=1}^{K} |\boldsymbol{H}(k,\theta)\boldsymbol{H}^H(k,\theta)\hat{\Phi}_{SS}(k) + \hat{\sigma}^2\boldsymbol{\Gamma}_{\boldsymbol{VV}}(k)|. \quad (47)$$

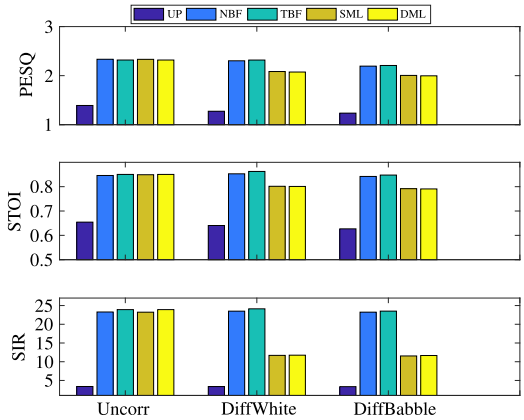


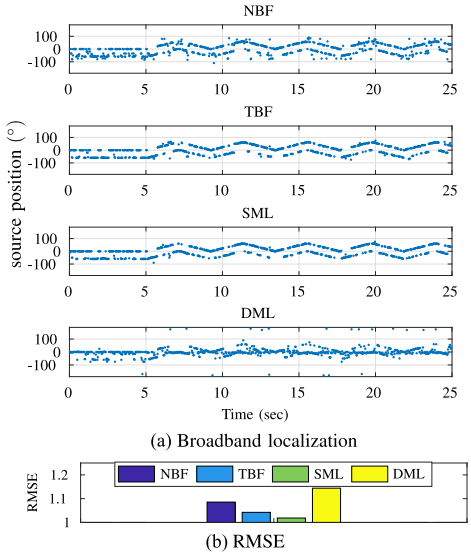
Fig. 5. The performance of the adaptive binaural beamformer for the separation of two active speakers at $\pm 30^\circ$ when the listener head is fixed. The experiments were conducted in a reverberant room with $T_{60} = 0.4$s with uncorrelated white noise (Uncorr), diffuse white noise (DiffWhite) and diffuse babble noise (DiffBabble). The global broadband SNR is 10 dB. The unprocessed signal (UP) has an SIR of 2 dB.

Since $\boldsymbol{H}(k,\theta)\boldsymbol{H}^H(k,\theta)$ is a rank-1 matrix, the determinant of the spatial covariance matrix is almost zero for high SNR bins. However, low SNR bins in which $\sigma^2\boldsymbol{\Gamma}_{\boldsymbol{VV}}$ is a full rank matrix have a non-zero determinant. Therefore, when the contributions of low SNR and high SNR bins are multiplied, reliable bands dominate the overall cost function and thus eliminate the effect of noisy bands. Hence, this leads to more precise broadband DOA estimates. In other approaches however, the contribution of reliable and noisy bins are averaged across all frequencies and thus the impact of high SNR bins will be less pronounced as compared to the SML algorithm.

We also evaluate the performance of the adaptive beamformer integrated with each of the four localization approaches and when the dummy head is in a fixed position. The performance is measured for the separation of two concurrent speakers located at $\pm 30^\circ$ w.r.t. the head. Results for the experiments in the reverberant room and under different noisy conditions with 10 dB SNR are reported in Fig. 5. According to this figure, the TBF approach outperforms other algorithms whereas the SML algorithm has less performance than other algorithms. It reveals that although the noise is characterized in the SML method, the narrowband DOA estimates are more sensitive to the mismatch of HRTFs and also to estimation errors in the spatial covariance of the noise as compared to beamforming-based localization algorithms. Also, the DML algorithm shows less improvement in diffuse noise as compared to the TBF approach which is due to the estimation errors in the noise covariance matrix.

B. Evaluation for Head Movements

In the next experiment we evaluate the performance of the proposed algorithms in a dynamic situation when the head of the listener turns. The two source loudspeakers are located at $\pm 30^\circ$ w.r.t. the head. One cycle of head turns starts when the dummy head is in front of the first speaker $(\theta_{S_1}, \theta_{S_2}) = (0^\circ, -60^\circ)$ and ends when the head is facing the second speaker $(\theta_{S_1}, \theta_{S_2}) = (60^\circ, 0^\circ)$. The angular speed of the head turn is $30^\circ/s$. The



(a) Broadband localization

(b) RMSE

Fig. 6. The broadband DOA estimates (a) and the RMSE (b) of two concurrent speakers when the listener's head turns. DOAs are estimated in the presence of diffuse babble noise with 10 dB SNR.

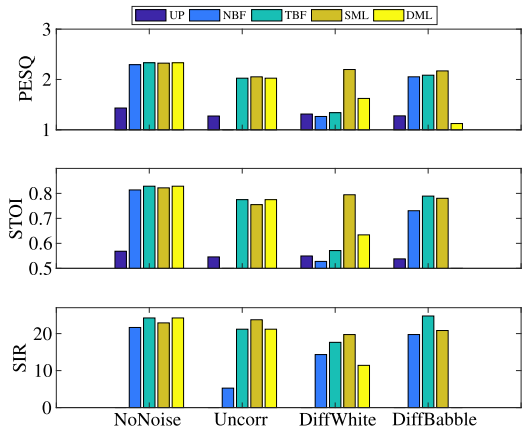


Fig. 7. The performance of the adaptive binaural beamformer for the separation of two active speakers at $\pm 30^\circ$ with the movement of the head with angular speed $30^\circ/s$. This experiment was done in a reverberant room with $T_{60} = 0.4$ s and without noise (NoNoise) and under different noisy conditions with uncorrelated white noise (Uncorr), diffuse white noise (DiffWhite) and diffuse babble noise (DiffBabble). The global broadband SNR is 10 dB. The unprocessed signal (UP) has an SIR of 0 dB.

broadband DOA estimates for the recording in the presence of diffuse babble noise with 10 dB SNR are displayed in Fig. 6. According to the example presented in Fig. 6(a), the SML algorithm has the best performance. Both the TBF and the SML algorithms are more robust as compared to the NBF and DML approaches which show outliers particularly for the lateral angles. Since the accuracy of the DML algorithm relies on the precise estimation of noise covariance, it achieves a low performance. Note that the broadband localization allows the estimation of only one source per frame. In order to analyze these results quantitatively we compute the root mean square error (RMSE) of all algorithms which is shown in Fig. 6(b). It verifies that the SML approach achieves less RMSE for the broadband localization in the presence of listener head turns as compared to the other algorithms.

1) Source Separation Using Online Broadband DOA Estimation: For static recordings we just use broadband DOA estimates to initialize the steering vector of the beamformer. However, when the head of the listener turns the corresponding steering vector changes rapidly. One solution is to track the movement of the listener head by means of a 9-axis inertial measurement unit (IMU) which was proposed in our previous works [32], [46]. The IMU provides the relative position of the head w.r.t. its initial position at each time step.

In the current work, however, we make use of the online broadband DOA estimates to adapt the steering vector of the binaural beamfomer. Since the broadband estimation provides one estimate per frame, we need to adapt the estimated DOAs of both target and interfering speakers. Therefore, we assign the new DOA estimate to a source if the estimate is in its vicinity. The vicinity threshold is set to $\pm 10^\circ$. When the DOA of one of the speakers is adapted we use the same update value to adapt the position of the other sources. This is possible since we only consider stationary source positions.

Results for different types of background noise with 10 dB SNR are summarized in Fig. 7. It is observed that in the absence of ambient noise the beamformers adapted with the four DOA algorithms attain similar performance. This verifies the ability of the proposed systems to localize, track and separate simultaneous speakers. Under the diffuse babble noise the TBF and SML approaches provide an improved performance as compared to others which is an expected result as shown in Fig. 6. In the presence of uncorrelated white noise the GSC combined with the SML algorithm outperforms other algorithms in terms of SIR, however, it delivers less predicted intelligibility than the other approaches. Additionally, for diffuse white noise the GSC integrated with the SML approach is superior to others since it relies on more precise broadband DOA estimates.

C. Influence of the Mismatch of HRTFs

We now investigate the effect of mismatch of HRTFs on the performance of the adaptive binaural beamformer. We consider two cases: In the first case binaural signals are generated using the HRTF database [27]. In the second case we record binaural signals in our anechoic room. In both cases speakers are located at $\pm 30^\circ$ w.r.t. the head and the same database [27] is used as HRTF prototypes in the binaural localization algorithms. Results are reported in Fig. 8. Since we consider a noise free environment the DML approach is identical to the TBF and is not shown here. The result verifies that the adaptive binaural beamformer is robust against the mismatch of HRTFs and achieves a reliable efficiency for the separation of two concurrent speakers. It also shows the robustness of all proposed DOA algorithms for the mismatch of HRTFs. Furthermore, we investigated in our recent work [47] the impact of the head radius on binaural localization. In certain applications with non-standard head sizes, e.g., for hearing-impaired children or robots, an adaptive binaural localization approach based on the joint optimization of head radius and DOA could be useful.

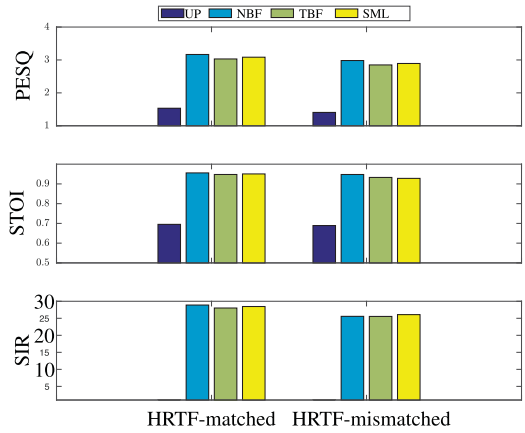


Fig. 8. Evaluation of the effect of HRTF mismatch on the adaptive binaural beamformer for the separation of two competing speakers located at $\pm 30^\circ$ w.r.t. the head. HRTF-matched and HRTF-mismatched indicate experiments in anechoic rooms tested with matched and mismatched HRTFs, respectively. The unprocessed signal (UP) has an SIR of 0 dB.

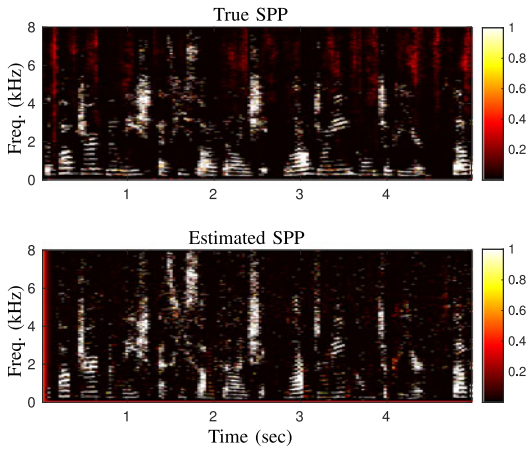


Fig. 9. Evaluation of the proposed SPP. True SPPs (top) obtained from the oracle information and the estimated SPPs (bottom) using SML and GSC.

D. Evaluation of the Proposed GSC Using SPP

We examine the accuracy of the estimated multichannel SPP in an anechoic room and in the presence of uncorrelated white noise with a global SNR of 10 dB. We use the HRIR database [27] to render binaural signals of two sources located at $\pm 30^\circ$. We use the SML algorithm to estimate the *a priori* SNR and use the power of the GSC output to estimate the *a posteriori* SNR. The result for one of the speakers is plotted in Fig. 9. It is observed that the estimated SPPs approximate the true values both in high and low narrowband SNRs which supports the benefit of using the SML algorithm.

Fig. 10 illustrates the performance of all approaches for the separation of two speakers in the reverberant room and under different background noise with 10 dB SNR. According to this Figure, the GSC-SPP approach achieves a better performance in terms of separation than other approaches. In terms of the quality and intelligibility the GSC-SPP method is superior compared to the original GSC integrated with SML and DML localization methods. However, the new approach shows a lower predicted intelligibility compared to the TBF-based GSC. This trade-off can be adjusted via the prior probabilities in the SPP equations.

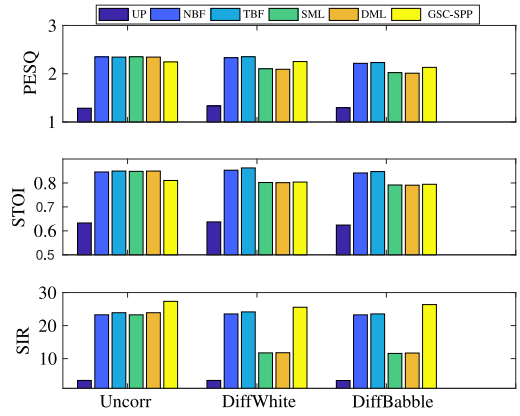


Fig. 10. Evaluation of proposed algorithms including GSC-SPP for the separation of two speakers at $\pm 30^\circ$. This experiment was done in a reverberant room with $T_{60} = 0.4$s with uncorrelated white noise (UncorrWhite), diffuse white noise (DiffWhite), and diffuse babble noise (DiffBabble). The global broadband SNR is 10 dB. The unprocessed signal (UP) has an SIR of 2 dB.

VIII. Discussion and Conclusion

In this work we presented a framework for binaural speaker localization using BTE HA microphones. The framework consists of two groups of algorithms either based on beamforming or ML techniques. The binaural cues used in our algorithms are characterized in the form of HRTFs. They are extracted from a database for BTE HAs with four microphones.

The beamforming-based localization algorithms steer the beam in a direction where either the maximum or the minimum response power is achieved. This results in the TBF and NBF methods, respectively. In both methods we filter the signal with the energy-normalized HRTF for all possible source locations. We also developed the ML-based binaural localization methods in which the background noise of the environment is modeled with a Gaussian distribution. Depending on how we treat the clean signal in the model we achieve two types of ML methods. The SML cost function needs the estimate of the noise power and its normalized spatial covariance whereas DML requires only the noise spatial covariance estimates.

We evaluated each of the proposed localization algorithms individually and for the task of speaker separation. Hence, we adapt the DOA-based GSC [4] to the binaural configuration. We use broadband DOA estimates to steer the beamformer towards the source and track the position of target and interfering sources while the listener's head may turn. The narrowband DOA estimates are then used to compute the TPP. It adapts the blocking matrix and the noise canceller of the GSC algorithm.

Results show that the SML approach outperforms other methods in terms of broadband DOA estimation especially in the presence of background noise and when the listener's head turns. However, the SML algorithm achieves less accuracy in narrowband DOA estimation since it is sensitive to estimation errors in the noise covariance and to the mismatch of HRTFs. However, in the broadband SML localization the contribution of all

frequency bins are multiplied. Therefore, the contribution of high SNR frequency bins dominate low SNR bins resulting in the most accurate DOA estimates. Although the noise is not modeled in the TBF approach, it performs practically promising in the static scenarios. The DML approach can be interpreted as an extension of the TBF approach considering spatial noise characteristics. It was shown for spatially uncorrelated noise that DML is identical to TBF. The performance of DML highly depends on the accuracy of noise covariance estimates. For diffuse noise this causes a larger error especially at lower SNR and low frequencies.

Theoretically, ML approaches are optimal for the DOA estimation provided that no mismatch of the acoustic condition and no estimation errors of noise statistics occur. Practically, the SML technique delivers a reliable performance especially in the presence of head movements. In addition, the SML approach estimates of the power of clean speech and noise signals. Therefore, we made use of them to estimate SPP and to develop a new adaptive beamformer. The new beamformer is able to model the noise as an additional Gaussian component in the estimation of the TPP. Then, the blocking matrix of the new beamformer is controlled by the incorporation of target presence and speech presence probabilities. This leads to an improved separation performance of the binaural beamformer.

In conclusion among all introduced localization algorithms, on the one hand, the TBF algorithm has a convincing performance for narrowband DOA estimation and source separation in static scenarios in addition to its simplicity for real-time implementation in HAs. On the other hand, despite the computational complexity of the SML algorithm, it offers more precise broadband DOA estimates than others and is well-suited for source tracking and separation under noisy conditions. Therefore, the proposed SPP-based adaptive beamformer derived from the SML localization is a good candidate for source separation in noisy and dynamic environment. The proposed localization methods deliver estimates for the full circle of azimuth angles since they use four HA microphones. Results show the robustness of the system against reverberation without requiring prior training of binaural cues.

APPENDIX
DERIVATION OF THE DML METHOD

In order to derive the DML localization cost function, we first obtain estimations of σ^2 and S and then replace them in (15) and optimize it with respect to θ. Taking the derivative of (15) w.r.t. σ^2 and setting it to zero yields

$$\frac{-BM}{\sigma^2} + \frac{1}{\sigma^4}\sum_{b=1}^{B}\left(\boldsymbol{X}(b) - \boldsymbol{H}S(b)\right)^H \boldsymbol{\Gamma}_{\boldsymbol{VV}}^{-1}\left(\boldsymbol{X}(b) - \boldsymbol{H}S(b)\right) = 0. \tag{48}$$

The estimate of the noise power is thus given by

$$\hat{\sigma}^2 = \frac{1}{BM}\sum_{b=1}^{B}\left(\boldsymbol{X}(b) - \boldsymbol{H}S(b)\right)^H \boldsymbol{\Gamma}_{\boldsymbol{VV}}^{-1}\left(\boldsymbol{X}(b) - \boldsymbol{H}S(b)\right). \tag{49}$$

Taking the derivative of (15) w.r.t. S and setting it to zero yields the clean signal

$$\hat{S} = \frac{\boldsymbol{H}^H\boldsymbol{\Gamma}_{\boldsymbol{VV}}^{-1}}{\boldsymbol{H}^H\boldsymbol{\Gamma}_{\boldsymbol{VV}}^{-1}\boldsymbol{H}}\boldsymbol{X}, \tag{50}$$

which is identical to the solution of the MVDR beamformer. Finally, by substituting (49) and (50) in (15) we obtain

$$L(\boldsymbol{X}^B|\theta, \boldsymbol{S}^B, \boldsymbol{\Phi}_{\boldsymbol{VV}}) = -BM\log\pi - B\log(|\boldsymbol{\Gamma}_{\boldsymbol{VV}}|) - BM\log\frac{1}{B}\sum_{b=1}^{B}(\boldsymbol{X} - \boldsymbol{H}S)^H\boldsymbol{\Gamma}_{\boldsymbol{VV}}^{-1}(\boldsymbol{X} - \boldsymbol{H}S) - BM, \tag{51}$$

which is maximized across θ. Therefore, we only consider the third term in (51) and drop the $\log$ function as it is monotonic. The DOA estimation is thus given by

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\left\{\frac{1}{B}\sum_{b=1}^{B}\left(\boldsymbol{X}(b) - \boldsymbol{H}S(b)\right)^H\boldsymbol{\Gamma}_{\boldsymbol{VV}}^{-1}\left(\boldsymbol{X}(b) - \boldsymbol{H}S(b)\right)\right\}. \tag{52}$$

We expand the argument in (52) to

$$\frac{1}{B}\sum_{b=1}^{B}\boldsymbol{X}^H(b)\boldsymbol{\Gamma}_{\boldsymbol{VV}}^{-1}\boldsymbol{X}(b) - \frac{1}{B}\sum_{b=1}^{B}\boldsymbol{X}^H(b)\boldsymbol{\Gamma}_{\boldsymbol{VV}}^{-1}\boldsymbol{H}S(b) - \frac{1}{B}\sum_{b=1}^{B}(\boldsymbol{H}S(b))^H\boldsymbol{\Gamma}_{\boldsymbol{VV}}^{-1}\boldsymbol{X}(b) + \frac{1}{B}\sum_{b=1}^{B}(\boldsymbol{H}S(b))^H\boldsymbol{\Gamma}_{\boldsymbol{VV}}^{-1}\boldsymbol{H}S(b). \tag{53}$$

If we insert (50) in (53) the third and the fourth term are equal and thus cancel. We may write

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\frac{1}{B}\sum_{b=1}^{B}\boldsymbol{X}^H(b)\boldsymbol{\Gamma}_{\boldsymbol{VV}}^{-1}\left(\boldsymbol{I}_{M\times M} - \boldsymbol{P}_{\boldsymbol{H}}(\theta)\right)\boldsymbol{X}(b), \tag{54}$$

where $\boldsymbol{P}_{\boldsymbol{H}}(\theta)$ is given by (17). We thus simplify (54) to

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\frac{1}{B}\sum_{b=1}^{B}\boldsymbol{X}^H(b)\boldsymbol{\Gamma}_{\boldsymbol{VV}}^{-1}\boldsymbol{P}_{\boldsymbol{H}}(\theta)\boldsymbol{X}(b). \tag{55}$$

ACKNOWLEDGMENT

The authors would like to thank Dr. Nilesh Madhu for his valuable comments concerning the DML localization algorithm. They would also like to thank the anonymous reviewers for their comments that have improved the clarity of this paper.

REFERENCES

[1] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Blind extraction of dominant target sources using ICA and time-frequency masking," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 2165–2173, Nov. 2006.

[2] K. Kumatani, J. McDonough, B. Rauch, D. Klakow, P. Garner, and W. Li, "Beamforming with a maximum negentropy criterion," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 5, pp. 994–1008, Jul. 2009.

[3] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 6, pp. 1071–1086, Aug. 2009.

[4] N. Madhu and R. Martin, "A versatile framework for speaker separation using a model-based speaker localization approach," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 1900–1912, Sep. 2011.

[5] H. Teutsch and G. W. Elko, "First-and second-order adaptive differential microphone arrays," in *Proc. Int. Workshop Acoust. Echo Noise Control*, 2001, pp. 35–38.

[6] N. Chatlani, E. Fischer, and J. J. Soraghan, "Spatial noise reduction in binaural hearing aids," in *Proc. 2010 18th Eur. Signal Process. Conf.*, Aug. 2010, pp. 1544–1548.

[7] H. Puder, E. Fischer, and J. Hain, "Optimized directional processing in hearing aids with integrated spatial noise reduction," in *Proc. Int. Workshop Acoust. Signal Enhanc.*, Sep. 2012, pp. 1–4.

[8] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*. New York, NY, USA: Wiley, 2006.

[9] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds. New York, NY, USA: Springer, 2001, pp. 39–60.

[10] S. Doclo and M. Moonen, "GSVD-based optimal filtering for multi-microphone speech enhancement," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds. New York, NY, USA: Springer, 2001, pp. 111–132.

[11] D. Marquardt, E. Hadad, S. Gannot, and 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/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 12, pp. 2384–2397, Dec. 2015.

[12] E. Hadad, S. Doclo, and S. Gannot, "The binaural LCMV beamformer and its performance analysis," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 3, pp. 543–558, Mar. 2016.

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

[14] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[15] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Trans. Signal Process.*, vol. 47, no. 10, pp. 2677–2683, Oct. 1999.

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

[17] M. Raspaud, H. Viste, and G. Evangelista, "Binaural source localization by joint estimation of ILD and ITD," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 1, pp. 68–77, Jan. 2010.

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

[19] J. Woodruff and D. Wang, "Binaural localization of multiple sources in reverberant and noisy environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 5, pp. 1503–1512, Jul. 2012.

[20] S. Goetze, T. Rohdenburg, V. Hohmann, B. Kollmeier, and K. D. Kammeyer, "Direction of arrival estimation based on the dual delay line approach for binaural hearing aid microphone arrays," in *Proc. 2007 Int. Symp. Intell. Signal Process. Commun. Syst.*, Nov. 2007, pp. 84–87.

[21] S. Braun, W. Zhou, and E. A. P. Habets, "Narrowband direction-of-arrival estimation for binaural hearing aids using relative transfer functions," in *Proc. 2015 IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2015, pp. 1–5.

[22] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 382–394, Feb. 2010.

[23] I. Merks, G. Enzner, and T. Zhang, "Sound source localization with binaural hearing aids using adaptive blind channel identification," in *Proc. 2013 IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 438–442.

[24] S. Mohan, M. Lockwood, M. Kramer, and D. Jones, "Localization of multiple acoustic sources with small arrays using a coherence test," *J. Acoust. Soc. Amer.*, vol. 123, no. 4, pp. 2136–2147, 2008.

[25] A. Archer-Boyd, W. Whitmer, W. Brimijoin, and J. Soraghan, "Biomimetic direction of arrival estimation for resolving front-back confusions in hearing aids," *J. Acoust. Soc. Amer.*, vol. 137, no. 5, pp. EL360–EL366, 2015.

[26] M. Farmani, M. Pedersen, Z. H. Tan, and J. Jensen, "Informed sound source localization using relative transfer functions for hearing aid applications," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 3, pp. 611–623, Mar. 2017.

[27] H. Kayser, S. D. Ewert, J. Anemüller, 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. Adv. Signal Process.*, vol. 2009, 2009, Art. no. 298605.

[28] M. Zohourian, G. Enzner, and R. Martin, "On the use of beamforming approaches for binaural speaker localization," in *Proc. 12. ITG Symp. Speech Commun.*, Oct. 2016, pp. 1–5.

[29] J. DiBiase, H. Silverman, and M. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds. New York, NY, USA: Springer, 2001, pp. 157–180.

[30] E. Jan, P. Svaizer, and J. L. Flanagan, "Matched-filter processing of microphone array for spatial volume selectivity," in *Proc. 1995 IEEE Int. Symp. Circuits Syst.*, vol. 2, Apr. 1995, pp. 1460–1463.

[31] N. Madhu and R. Martin, "Acoustic source localization with microphone arrays," in *Advances in Digital Speech Transmission*, R. Martin, U. Heute, and C. Antweiler, Eds. New York, NY, USA: Wiley, 2008.

[32] M. Zohourian and R. Martin, "Binaural speaker localization and separation based on a joint ITD/ILD model and head movement tracking," in *Proc. 2016 IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2016, pp. 430–434.

[33] J. F. Böhme and D. Kraus, "On least squares methods for direction of arrival estimation in the presence of unknown noise fields," in *Proc. 1998 IEEE Int. Conf. Acoust., Speech Signal Process.*, 1988, pp. 2833–2836.

[34] H. Ye and R. DeGroat, "Maximum likelihood DOA estimation and asymptotic Cramér-Rao bounds for additive unknown colored noise," *IEEE Trans. Signal Process.*, vol. 43, no. 4, pp. 938–949, Apr. 1995.

[35] T. Lotter and P. Vary, "Dual-channel speech enhancement by superdirective beamforming," *EURASIP J. Appl. Signal Process.*, vol. 2006, pp. 175–175, 2006.

[36] S. Araki, T. Nakatani, H. Sawada, and S. Makino, "Blind sparse source separation for unknown number of sources using Gaussian mixture model fitting with Dirichlet prior," in *Proc. 2009 IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2009, pp. 33–36.

[37] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[38] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc.*, vol. 39, pp. 1–38, 1977.

[39] M. Taseska and E. A. P. Habets, "An online EM algorithm for source extraction using distributed microphone arrays," in *Proc. 21st Eur. Signal Process. Conf.*, Sep. 2013, pp. 1–5.

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

[41] M. Souden, J. Chen, J. Benesty, and S. Affes, "Gaussian model-based multichannel speech presence probability," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 5, pp. 1072–1077, Jul. 2010.

[42] P. Kabal, "TSP speech database," McGill Univ., Montreal, QC, Canada, Database Version: 01, 2002.

[43] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-A new method for speech quality assessment of telephone networks and codecs," in *Proc. 2001 IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, 2001, pp. 749–752.

[44] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[45] C. Févotte *et al.*, "BSS_EVAL toolbox user guide–Revision 2.0," IRISA, Rennes, France, Tech. Rep. 1706, 2005.

[46] M. Zohourian, A. Archer-Boyd, and R. Martin, "Multi-channel speaker localization and separation using a model-based GSC and an inertial measurement unit," in *Proc. 2015 IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2015, pp. 5615–5619.

[47] M. Zohourian, R. Martin, and N. Madhu, "New insights into the role of the head radius in model-based binaural speaker localization," in *Proc. 2017 25th Eur. Signal Process. Conf.*, Aug. 2017, pp. 221–225.

Mehdi Zohourian (S'15) received the B.Sc. degree from Isfahan University of Technology, Isfahan, Iran, in 2010, and the M.Sc. degree from Sharif University of Technology, Tehran, Iran, in 2012, both in electrical engineering. Since 2014, he has been working toward the Ph.D. degree at the Institute of Communication Acoustics, Ruhr-Universität Bochum, Bochum, Germany. He was a Marie Curie Fellow of Improved Communication through applied hearing research (ICanHear) training network from 2014 to 2016. His research interests include the area of statistical signal processing with more emphasis on acoustic source localization and beamforming, multichannel speech enhancement, and hearing aids.

Gerald Enzner (S'00–M'06–SM'12) received the Dipl.-Ing. degree from the University of Erlangen Nürnberg, Erlangen, Germany, in 2000, and the Dr.-Ing. degree from RWTH Aachen University, Aachen, Germany, in 2006, both in electrical engineering.

Since 2007, he has been working as a Principal Investigator with the Institute of Communication Acoustics, Ruhr-Universität Bochum, Bochum, Germany, in manifold cooperations with public and private funding organizations. He was a Member of the Global Young Faculty with the University Alliance Metropolis Ruhr and the Mercator Foundation in the 2009–2011 term. He then was appointed to faculty level via Habilitation (2013) and Inauguration (2016) with the Department of Electrical Engineering and Information Technology, Ruhr-Universität Bochum. His research interests include statistical signal processing, single- and multichannel adaptive filtering, nonlinear adaptive systems, time-varying system models and algorithms, and blind channel identification and equalization. The research is currently applied in audio and acoustic signal processing, such as for acoustic echo and noise control in telecommunications, speech dereverberation and denoising, and signal processing for spatial sound control.

Dr. Enzner has been serving as the Vice-Chair of the Germany Chapter of the IEEE Signal Processing Society, since 2016.

Rainer Martin (S'86–M'90–SM'01–F'11) received the M.S.E.E. degree from the Georgia Institute of Technology, Atlanta, GA, USA, in 1989, and the Dipl.-Ing. and Dr.-Ing. degrees from RWTH Aachen University, Aachen, Germany, in 1988 and 1996, respectively.

From 1996 to 2002, he was a Senior Research Engineer with the Institute of Communication Systems and Data Processing, RWTH Aachen University. From April 1998 to March 1999, he was a Technology Consultant with the AT&T Speech and Image Processing Services Research Lab (Shannon Labs), Florham Park, NJ. From April 2002 until October 2003, he was a Professor of digital signal processing with the Technische Universität Braunschweig, Braunschweig, Germany. Since October 2003, he has been a Professor of information technology and communication acoustics with Ruhr-Universität Bochum, Bochum, Germany, and from October 2007 to September 2009, the Dean with the Electrical Engineering and Information Sciences Department. He is the Co-Author with P. Vary of *Digital Speech Transmission Enhancement, Coding and Error Concealment* (Wiley, 2006) and the Co-Editor with U. Heute and C. Antweiler of *Advances in Digital Speech Transmission* (Wiley, 2008). His research interests include signal processing for voice communication systems, hearing instruments, and humanmachine interfaces. He served as an Associate Editor for the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING and is a Member of the Speech and Language Processing Technical Committee of the IEEE Signal Processing Society.