GSC-BASED BINAURAL SPEAKER SEPARATION PRESERVING SPATIAL CUES

Mehdi Zohourian, and Rainer Martin

Institute of Communication Acoustics
Ruhr-Universität Bochum, Germany
{mehdi.zohourian, rainer.martin}@rub.de

ABSTRACT

In this paper we investigate two methods for the preservation of spatial cues in binaural speaker separation. We develop these methods as extensions of our previously proposed model-based generalized sidelobe canceller (GSC) which utilizes a maximum likelihood technique for speaker localization. In the proposed implementation the adaptive GSC provides an estimate of the target signal as well as an estimation of target presence probability (TPP). Binaural outputs are generated in two different ways: In the first approach the binaural signals are rendered using the GSC output signal combined with the HRTF hypotheses which are adapted by the broadband localization. The second approach uses the GSC output and the TPP to determine a common spectral postfilter. We find that the adaptive beamformer combined with the binaural rendering technique leads to larger improvements of the quality of the desired signal and delivers less unnatural fluctuations as compared to the common spectral postfilter. Informal subjective tests as well as instrumental measurements in the presence of the listener head movements reveals, however, the benefit of the spatially motivated spectral postfilter for the preservation of binaural cues of both target and interferer signals.

Index Terms— Binaural source localization, beamforming, source separation, hearing aids, generalized sidelobe canceller

1. INTRODUCTION

Speech enhancement in hearing aids (HAs) still provides many challenges especially in realistic and divers acoustic scenarios. However, the advent of the wireless link between the left and right HAs allows the execution of multichannel algorithms that result in better speech quality and intelligibility [1] as well as reliable sound source localization as compared to monaural processing.

Several studies have investigated the problem of the attenuation of directional interferes such as competing speakers. These may be divided into three main categories. The first group is the adaptive differential microphone array and its extensions [2, 3] that suppress directional noise without having prior knowledge of source positions. The second group of algorithms, e.g., [4, 5, 6] exploit directional probabilistic models to build time-frequency masks in order to segregate speech sources. The third group of algorithms are beamforming techniques that are commonly combined with localization methods to minimize the mismatch of the source direction and the steering vector [7, 8].

One of the main objectives in binaural speech enhancement is to preserve the binaural cues, namely interaural level difference (ILD) and the interaural time/phase difference (ITD/IPD). Various techniques have been proposed to deal with the problem of binaural noise reduction preserving these spatial cues. These can be classified in two main groups. Some authors have developed adaptive beamforming algorithms taking additional constraints into accounts that balances between noise reduction and cue preservation [9, 10, 11]. Other studies have applied a real-valued common gain to both left and right channels [12, 5, 13].

In this study we aim at the separation of simultaneous speakers while preserving the binaural cues of the input signal. We use *behind-the-ear* (BTE) HAs with 2x2 microphones. We extend our previously proposed model-based GSC [14] that employs the stochastic maximum likelihood (SML) localization technique [15]. The SML also provides the estimate of the power of the clean speech and noise signals. These estimates are utilized to compute the target presence probability (TPP). Then, we generate binaural outputs in two different ways: The first approach is to render the binaural signals using the GSC output signal in conjunction with the head-related transfer function (HRTF) hypotheses that are estimated based the localization results. The other approach derives the spectral postfilter motivated by GSC output and the TPP estimate.

The remainder of this paper is organized as follows. Section 2 introduces the binaural signal model. In Section 3 the binaural localization using the SML approach is described. Section 4 and 5 discuss the TPP estimation and the model-based GSC, respectively. In Section 6 we propose two binaural cue preservation approaches. Experimental results are explained in Section 7. Section 8 concludes this paper.

2. BINAURAL SIGNAL MODEL

We consider multichannel signals from Q sources received by the M microphones of binaural HAs. Analyzing signals in the STFT domain and using matrix notation we obtain the received signal as

$$\boldsymbol{X}(k,b) = \boldsymbol{H}(k,\boldsymbol{\Theta})\boldsymbol{S}(k,b) + \boldsymbol{V}(k,b), \tag{1}$$

where (k, b) indicate frequency and frame indices. $\boldsymbol{S}$ and $\boldsymbol{V}$ represent spectral vectors of the point source signal and the noise signal at microphones, respectively. In principle, $\boldsymbol{H}$ is the matrix of binaural room transfer functions (BRTFs) of the M microphones determined by

$$\boldsymbol{H}(k,\boldsymbol{\Theta}) = [\boldsymbol{H}_1(k,\theta_1), \boldsymbol{H}_2(k,\theta_2), .., \boldsymbol{H}_Q(k,\theta_Q)], \tag{2}$$

where $\boldsymbol{H}_q(k,\theta_q) = [H_{q1}(k,\theta_q), H_{q2}(k,\theta_q), .., H_{qM}(k,\theta_q)]^T$. In this equation θ_q denotes the azimuth location of source q. The signal vectors are given by $\boldsymbol{X}(k,b) = [X_1(k,b), X_2(k,b), .., X_M(k,b)]^T$, $\boldsymbol{S}(k,b) = [S_1(k,b), S_2(k,b), .., S_Q(k,b)]^T$, and $\boldsymbol{V}(k,b) = [V_1(k,b), V_2(k,b), .., V_M(k,b)]^T$. The received signals are processed by the GSC which is adapted using the SML DOA estimation approach [15].

This work has received funding from the European Fund for Regional Development, grant no EFRE-0800372 (RaVis-3D).

3. SML LOCALIZATION ALGORITHM

If we assume that the source and the noise signals are short-time stationary stochastic random processes and both follow complex Gaussian distributions, we may write the probability density function of the narrowband signal $\boldsymbol{X}$ in each frequency bin as [15]

$$P(\boldsymbol{X}|\theta,\boldsymbol{\Phi_{XX}}) = \frac{1}{\pi^M|\boldsymbol{\Phi_{XX}}|}\exp\left(-\boldsymbol{X}^H\boldsymbol{\Phi}_{\boldsymbol{XX}}^{-1}\boldsymbol{X}\right), \quad (3)$$

where $\Phi_{\boldsymbol{XX}}$ is the spatial covariance matrix. We assume that a sequence of DFT frames of the narrowband signal $\boldsymbol{X}^B = [\boldsymbol{X}(1),...,\boldsymbol{X}(b),...,\boldsymbol{X}(B)]^T$ is temporarily independent and identically distributed. We thus write the log-likelihood function as

$$L(\boldsymbol{X}^B|\theta,\boldsymbol{\Phi_{XX}}) = \log\prod_{b=1}^{B} P(\boldsymbol{X}(b)|\theta,\boldsymbol{\Phi_{XX}}) \quad (4)$$
$$= -BM\log\pi - B\log|\boldsymbol{\Phi_{XX}}| - B\,\mathrm{Tr}\left\{\boldsymbol{\Phi}_{\boldsymbol{XX}}^{-1}\hat{\boldsymbol{\Phi}}_{\boldsymbol{XX}}\right\},$$

where $\hat{\Phi}_{\boldsymbol{XX}}$ is the estimation of the spatial covariance matrix using non-recursive averaging. Based on the signal model in (1) and assuming the spectral disjointness property of speech signals and a homogeneous noise field, we may write the spatial covariance matrix of the microphone signals as

$$\boldsymbol{\Phi_{XX}}(k) = \boldsymbol{H}(k,\theta)\boldsymbol{H}^H(k,\theta)\Phi_{SS}(k) + \Phi_{\boldsymbol{VV}}(k)\boldsymbol{\Gamma_{VV}}(k), \quad (5)$$

where Φ_{SS} and $\Phi_{\boldsymbol{VV}}$ denote the power of the dominant clean speech and the noise signals, respectively. Here, $\boldsymbol{\Gamma_{VV}}$ indicates the normalized spatial covariance matrix of the noise signal. The SML approach first derives the estimation of the power of the clean speech and noise signals $\hat{\Phi}_{SS}$ and $\hat{\Phi}_{\boldsymbol{VV}}$ conditioned on the DOA parameter θ. Then, it substitutes the estimated parameters in the log-likelihood function (4) and maximize it with respect to θ. The power of the clean signal is estimated by [15]

$$\hat{\Phi}_{SS}(k) = \quad (6)$$
$$\frac{\boldsymbol{H}^H(k,\theta){\boldsymbol{\Gamma_{VV}}}^{-1}(k)(\hat{\boldsymbol{\Phi}}_{\boldsymbol{XX}}(k) - \hat{\Phi}_{\boldsymbol{VV}}\boldsymbol{\Gamma_{VV}}(k))\boldsymbol{\Gamma}_{\boldsymbol{VV}}^{-1}(k)\boldsymbol{H}(k,\theta)}{\left(\boldsymbol{H}^H(k,\theta)\boldsymbol{\Gamma}_{\boldsymbol{VV}}^{-1}(k)\boldsymbol{H}(k,\theta)\right)^2}.$$

The estimation of the power of the noise signal is obtained as [15]

$$\hat{\Phi}_{\boldsymbol{VV}}(k) = \frac{1}{M-Q}\mathrm{Tr}\left(\boldsymbol{P}_{\boldsymbol{H}}^{\perp}(k,\theta)\hat{\boldsymbol{\Phi}}_{\boldsymbol{XX}}(k)\boldsymbol{\Gamma}_{\boldsymbol{VV}}^{-1}(k)\right). \quad (7)$$

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

$$\boldsymbol{P_H}(k,\theta) = \quad (8)$$
$$\boldsymbol{H}(k,\theta)\left(\boldsymbol{H}^H(k,\theta)\boldsymbol{\Gamma}_{\boldsymbol{VV}}^{-1}(k)\boldsymbol{H}(k,\theta)\right)^{-1}\boldsymbol{H}^H(k,\theta)\boldsymbol{\Gamma}_{\boldsymbol{VV}}^{-1}(k),$$

has the properties of an orthogonal projection matrix and $\boldsymbol{I}_{M\times M}$ is the identity matrix. Therefore, by substituting (6), (7) in (5) and inserting (5) in (4) we achieve the DOA cost function as [15]

$$\Lambda_{SML}(k,\theta) = \quad (9)$$
$$-\log|\boldsymbol{P_H}(k,\theta)\hat{\boldsymbol{\Phi}}_{\boldsymbol{XX}}(k)\boldsymbol{P}_{\boldsymbol{H}}^H(k,\theta) + \hat{\Phi}_{\boldsymbol{VV}}\boldsymbol{P}_{\boldsymbol{H}}^{\perp}(k,\theta)\boldsymbol{\Gamma_{VV}}(k)|,$$

which is maximized across all azimuth candidates using a grid search in steps of 5 degree. This provides narrowband DOA estimates $\hat{\theta}(k,b)$. We also sum the cost function across all frequency bins and take the global maximum and discard other maxima to achieve the broadband localization result. It improves the robustness of DOA estimates but delivers a single directional estimate per time frame only [16]. Note that, in the above equation we use HRTFs hypotheses $\hat{\boldsymbol{H}}(k,\theta)$ extracted from the database [17] instead of $\boldsymbol{H}(k,\theta)$.

4. TARGET PRESENCE PROBABILITY (TPP)

Due to the assumption that speech signals are spectrally disjoint, each time-frequency bin is dominated by one speech source. Given the DOA of Q concurrent speakers in the presence of ambient noise we may statistically formulate $Q+1$ hypotheses

- $\mathcal{H}_{S_q}(k,b), q\in\{1,..Q\}$, speech source q is present,
- $\mathcal{H}_V(k,b)$, all speech sources are absent.

The narrowband DOA estimates $\hat{\theta}(k,b)$ are commonly modeled using a Gaussian mixture model (GMM) [6, 7]. Therefore, the posterior probability of each hypothesis given the DOA estimates, which is known as TPP, is determined by

$$p(\mathcal{H}_{S_q}|\hat{\theta}(k,b)) = \frac{\rho_{S_q}\mathcal{N}\left(\hat{\theta}(k,b)|\mu_{S_q},\sigma_{S_q}^2\right)}{\sum_{i=1}^{Q+1}\rho_{S_i}\mathcal{N}\left(\hat{\theta}(k,b)|\mu_{S_i},\sigma_{S_i}^2\right)}, \quad (10)$$

where the mean of each Gaussian component indicates the location of each source, i.e., $\theta_{S_q} = \mu_{S_q}$. In this equation, $\sigma_{S_q}^2$ and ρ_{S_q} denote the variance and the prior probability of the q-th Gaussian component. The noise is also modeled as an extra Gaussian component in the GMM. The mean and the variance of this Gaussian distribution is set to $\mu_N = \pi$ and $\sigma_N^2 = \pi$. We set the variance of the speakers positions to $\sigma_{S_q}^2 = \pi/18$. Then, the priors ρ_{S_q} of the GMM are estimated using the multichannel speech presence probability (SPP) [18] in each frequency bin. We compute the conditional probability of each hypothesis given the noisy signal as [19]

$$p(\mathcal{H}_{S_q}|X) = \frac{p(X|\mathcal{H}_{S_q})p(\mathcal{H}_{S_q})}{\sum_{i=1}^{Q}p(X|\mathcal{H}_{S_i})p(\mathcal{H}_{S_i}) + p(X|\mathcal{H}_V)p(\mathcal{H}_V)}, \quad (11)$$

in which $p(\mathcal{H}_{S_q})$ and $p(\mathcal{H}_V)$ denote prior probabilities of the q-th speech source presence and speech absence, respectively. Note that in this equation we omit (k,b) for readability. Equation (11) may be simplified to $p(\mathcal{H}_{S_q}|X) = \frac{\Lambda_q}{1+\sum_{i=1}^{Q}\Lambda_i}$, where

$$\Lambda_q = \frac{p(\mathcal{H}_{S_q})}{p(\mathcal{H}_V)}\frac{p(X|\mathcal{H}_{S_q})}{p(X|\mathcal{H}_V)} \quad (12)$$

is the generalized likelihood ratio for source q. Assuming that DFT coefficients of speech and noise signals follow complex Gaussian distributions and are mutually independent we obtain the generalized likelihood ratio as [14]

$$\Lambda_q = \frac{p(\mathcal{H}_{S_q})}{p(\mathcal{H}_V)}\frac{1}{1+\delta_q\zeta_q}\exp\left(\frac{\delta_q^2\zeta_q}{1+\delta_q\zeta_q}\gamma_q\right), \quad (13)$$

where $\zeta_q = \frac{\Phi_{S_qS_q}}{\Phi_{\boldsymbol{VV}}}$ and $\gamma_q = \frac{\Phi_{\hat{S}_q\hat{S}_q}}{\Phi_{\boldsymbol{VV}}}$ are the *a priori* and the *a posteriori* SNR of source q respectively. Here, $\Phi_{S_qS_q}$ and $\Phi_{\boldsymbol{VV}}$ are estimated using (6) and (7), respectively. Moreover, $\Phi_{\hat{S}_q\hat{S}_q}$ denotes the power of the estimated source signal, where $\hat{S}_q$ is the GSC output given by (18). In (13), we have $\delta_q = \boldsymbol{H}^H(k,\theta_q)\boldsymbol{\Gamma}_{\boldsymbol{VV}}^{-1}\boldsymbol{H}(k,\theta_q)$. Under the special case of uncorrelated white noise using a free-field microphone array we have $\delta_q = M$. Therefore, for $M=1$ we achieve the well-known expression for the generalized likelihood ratio for single channel noise reduction [20]. We use (11) to estimate ρ_{S_q} in (10). Note that we assume fixed values for the prior probabilities of the q-th speech source presence and speech absence in (11). The TPP estimates are integrated in the model-based GSC [7, 14].

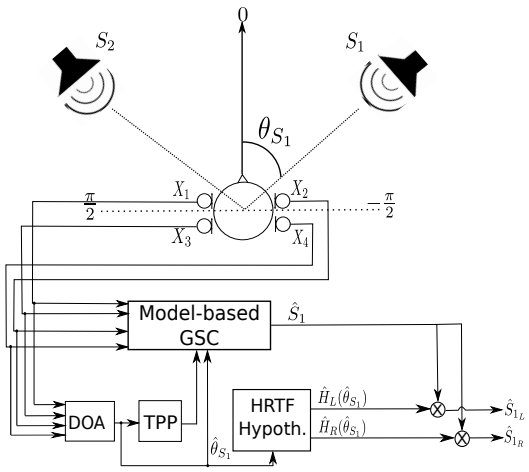


Fig. 1. Block diagram of the proposed binaural speaker separation algorithm using the GSC output signal combined with binaural rendering approach.

5. MODEL-BASED GSC

The GSC for source q consists of a fixed beamformer $\boldsymbol{W}_{f_q}(k,b)$, an adaptive blocking matrix $\boldsymbol{B}_q(k,b)$, and an adaptive noise canceler $\boldsymbol{W}_{V_q}(k,b)$. We design the fixed beamformer using the MVDR approach assuming uncorrelated noise denoted as $\boldsymbol{W}_{f_q}(k,\theta_q) = \frac{\boldsymbol{H}(k,\theta_q)}{||\boldsymbol{H}(k,\theta_q)||^2}$. It thus involves both IPD and ILD cues in its computation.

We use the same strategy as in [7, 8] for the design of the adaptive blocking matrix and its application to the binaural configuration. The idea is to first construct the projection matrix onto the target signal subspace which is estimated by

$$\hat{\boldsymbol{P}}_q(k,b) = \left(1 - p(\mathcal{H}_{S_q}|\hat{\theta}(k,b)\right)\hat{\boldsymbol{P}}_q(k,b-1) + p(\mathcal{H}_{S_q}|\hat{\theta}(k,b))\frac{\hat{\boldsymbol{\Phi}}_{\boldsymbol{XX}}(k,b)}{\|\mathbf{X}(k,b)\|^2} \quad , \tag{14}$$

and then to compute the projection to the complementary subspace as

$$\hat{\boldsymbol{P}}_q^{\perp}(k,b) = \boldsymbol{I}_{M\times M} - \hat{\boldsymbol{P}}_q(k,b). \tag{15}$$

Then, the blocking matrix is given by selecting the first $(M-1)$ rows and M columns of the matrix argument using an operator $\kappa_{(M-1)M}(\cdot)$ as

$$\boldsymbol{B}_q(k,b) = \kappa_{(M-1)M}\left(\hat{\boldsymbol{P}}_q^{\perp}(k,b)\right). \tag{16}$$

The adaptive noise canceller uses a normalized least mean-square (NLMS) algorithm [7]

$$\boldsymbol{W}_{V_q}(k,b+1) = \boldsymbol{W}_{V_q}(k,b) + \alpha_q \frac{\hat{S}_q^*(k,b)\boldsymbol{B}_q(k,b)\boldsymbol{X}(k,b)}{||\boldsymbol{B}_q(k,b)\boldsymbol{X}(k,b)||^2}, \tag{17}$$

with an adaptive step-size $\alpha_q = \left(1 - p(\mathcal{H}_{S_q}|\hat{\theta}(k,b))\right)\alpha_f$, where α_f denotes a fixed stepsize factor. In this equation, $\hat{S}_q$ indicates the output of the GSC beamformer for source q determined by

$$\hat{S}_q = \left(\boldsymbol{W}_{f_q}^H(k,b) - \boldsymbol{W}_{V_q}^H(k,b)\boldsymbol{B}_q(k,b)\right)\boldsymbol{X}(k,b). \tag{18}$$

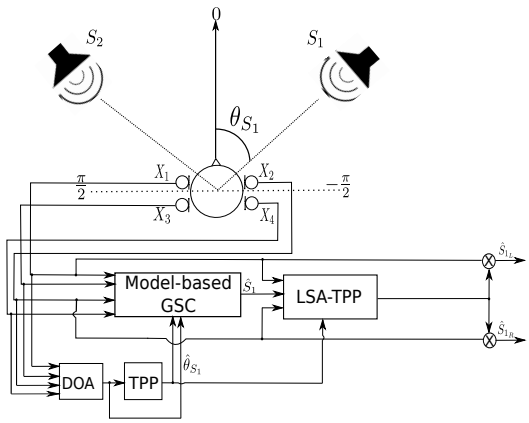


Fig. 2. Block diagram of the proposed binaural speaker separation algorithm using the GSC output signal for the estimation of the LSA-TPP postfilter.

6. BINAURAL CUE PRESERVATION

A fundamental step in binaural speech enhancement is to generate output signals that preserves spatial information of the original signal [21]. We consider two methods to generate dual-channel output from the beamformer. The first proposal is to render binaural signals using the localization results in conjunction with the HRTF hypotheses. We use the online broadband DOA estimation provided by the SML algorithm in order to determine the location of the target source and to adapt the corresponding HRTF hypothesis at each time step. We then multiply the GSC output signal in the frequency domain with the HRTFs hypothesis for the left and the right ear. The hypothesis could be extracted either from a database [17] or from the spherical head model [22]. The overall structure is shown in Fig. 1. This approach is similar to the binaural MVDR beamformer [11] method that applies a distortionless response constraint for the left and the right channel, separately.

An alternative is to utilize the beamformer output in a postfilter as a spectral gain [12]. In theory if we had a perfect estimate of the power of clean speech and noise signal, the optimal solution to the noise reduction problem would be the multichannel Wiener filter which is decomposed into a MVDR beamformer followed by a single channel Wiener filter. In practice since these estimates are erroneous, the Wiener gain would be a suboptimal solution for the postfilter. Furthermore, the proposed adaptive beamformer provides an additional flexibility to incorporate the TPPs in the single channel noise reduction.

In [23] a modified minimum mean-square error log-spectral amplitude (MM-LSA) estimator is proposed in which the SPP is multiplied by a spectral gain which is derived using the MMSE-LSA algorithm [24]. We use the same strategy for the design of the postfilter which is presented in Fig. 2. The spectral gain of the postfilter is denoted by

$$G_{LSA-TPP} = p(\mathcal{H}_{S_q}|\hat{\theta}(k,b))\frac{M\Phi_{\hat{S}_q\hat{S}_q}(k)}{\mathrm{Tr}\{\Phi_{XX}(k)\}}\exp\left(\frac{1}{2}\int_{\nu_q}^{\infty}\frac{e^{-t}}{t}dt\right), \tag{19}$$

where $\nu_q = \frac{M\Phi_{\hat{S}_q\hat{S}_q}(k)}{\mathrm{Tr}\{\Phi_{XX}(k)\}}\gamma_q$. The common gain is applied to the front left and right microphones of hearing aid to generate binaural outputs.

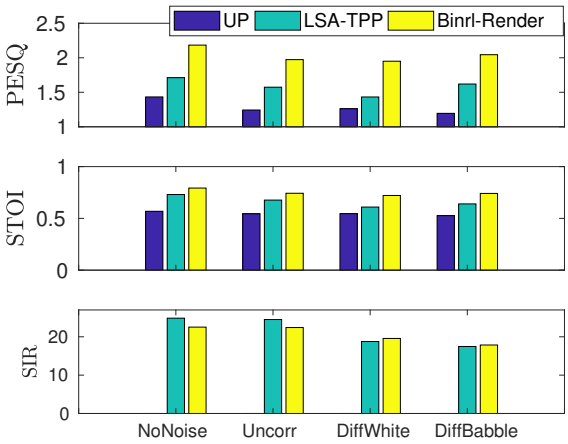


Fig. 3. Objective evaluation of proposed binaural speaker separation approaches for the recording with the listener head turns. Signals are recorded under no background noise (NoNoise) and under uncorrelated white noise (Uncorr), diffuse white noise (DiffWhite) and diffuse babble noise (DiffBabble) with global broadband SNR of 10 dB. The unprocessed signal (UP) has an SIR of 0 dB.

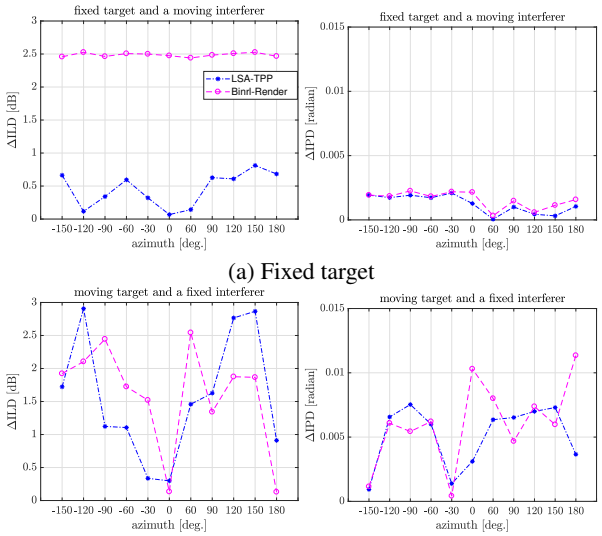
7. EVALUATION RESULTS

We conduct experiments in a reverberant room with ($T_{60} = 0.4$s, critical distance 1.1 m). We use the front and back microphones of a pair of BTE HAs attached to a dummy head. Loudspeakers playing male and female utterances are placed 1.2 m away from the dummy head and are thus outside the critical distance. We use speech signals from the TSP database [25]. Each signal consists of four male and four female utterances of 10 s duration. 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 FFT bins equals 1024.

In the first experiment we assess the performance of the proposed binaural speaker separation algorithms for a recording with listener head turns. 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 and ends when the head is facing the second speaker. The angular speed of the head turn is $30^\circ/s$.

We evaluate the performance of algorithms in terms of perceptual evaluation of speech quality (PESQ) [26], the short-time objective intelligibility (STOI) [27], and signal-to-interference ratio (SIR) [28]. Results for different types of background noise including uncorrelated noise, spatially diffuse white noise and spatially diffuse babble noise with 10 dB SNR are reported in Fig. 3. For this evaluation we use the front left microphone of hearing aids. Additionally, the motion of the head is tracked via the SML broadband localization. As it is observed from this figure the binaural rendering (Binrl-Render) approach outperforms the LSA-TPP approach in terms of the predicted quality and intelligibility. However, the LSA-TPP approach achieves a better result for the separation of the target signal. Furthermore, the binaural rendering approach shows less musical noise and thus fewer distortion than the LSA-TPP algorithm.

We also evaluate the capability of the proposed algorithms to preserve spatial cues. In the first experiment we consider the fixed target position at 30° and the moving interferer at locations in the full azimuth circle starting from -150° and increasing clockwisely with steps of 30° w.r.t. the head. In the second experiment the target is located at the aforementioned angles and the interferer angle is



(a) Fixed target

(b) Moving target

Fig. 4. ILD and IPD error of the proposed binaural speaker separation algorithms: (a) The target speaker is fixed at 30° and an interferer is at corresponding angles. (b) The target speaker is located at corresponding angles while the interferer is fixed at 30° .

set to 30°. Results in terms of ILD and IPD errors are presented in Fig. 4. Figure 4 (a) verifies that the binaural cues of the target signal using the LSA-TPP approach are better preserved than by using the binaural rendering approach when the target is fixed and the interferer moves. The reason lies in the mismatch of HRTFs as well as the effect of room reverberation that distorts the spatial cues in the binaural rendering approach. Moreover, as shown in Fig. 4 (b) the preservation of binaural cues varies depending on where the target is located. For angles in the frontal hemisphere the LSA-TPP approach outperforms the binaural rendering one. However, for angles in the back hemisphere the LSA-TPP shows less accuracy in the preservation of ILD cues than the binaural rendering approach. Nevertheless, there is only a small difference between the two approaches in terms of IPD error.

Based on informal listening test for the recording with head movements, the LSA-TPP approach is able to preserve the spatial cues of the residual interference, since the binaural rendering approach aligns the spatial cues of the residual interferer with the target.

8. CONCLUSION

In this paper we develop novel algorithms for the preservation of spatial cues to binaural speaker separation using HA microphones. We first combine the model-based GSC with stochastic maximum likelihood localization. The proposed structure delivers the target signal estimation in addition to an estimate of the target presence probability. The adaptive beamformer is then combined with two approaches to generate binaural outputs. The first approach is based on a binaural rendering technique using the GSC output signal in conjunction with HRTFs hypothesis derived from the localization. The other algorithm computes a common spectral gain which is applied to the left and right channels. Results corroborate that the LSA-TPP approach achieves better performance for the suppression of interfering signals and preservation of binaural cues of both target and interferer signals as compared to the binaural rendering method, however at the expense of more level of random fluctuations in the output signal.

9. REFERENCES

[1] H. Luts, K. Eneman, J. Wouters, M. Schulte, M. Vormann, M. Büchler, N. Dillier, R. Houben, W. A. Dreschler, M. Froehlich, et al., "Multicenter evaluation of signal enhancement algorithms for hearing aids," *The Journal of the Acoustical Society of America*, vol. 127, no. 3, pp. 1491–1505, 2010.

[2] H. Teutsch and G. W. Elko, "First-and second-order adaptive differential microphone arrays," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*. Citeseer, 2001, pp. 35–38.

[3] H. Puder, E. Fischer, and J. Hain, "Optimized directional processing in hearing aids with integrated spatial noise reduction," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2012, pp. 1–4.

[4] N. Roman, D. Wang, and Guy J. Brown, "Speech segregation based on sound localization," *The Journal of the Acoustical Society of America*, vol. 114, no. 4, pp. 2236–2252, 2003.

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

[6] 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. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2009, pp. 33–36.

[7] N. Madhu and R. Martin, "A versatile framework for speaker separation using a model-based speaker localization approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 1900–1912, Sept 2011.

[8] M. Zohourian and R. Martin, "Binaural speaker localization and separation based on a joint ITD/ILD model and head movement tracking," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2016, pp. 430–434.

[9] T. Van den Bogaert, J. Wouters, S. Doclo, and M. Moonen, "Binaural cue preservation for hearing aids using an interaural transfer function multichannel Wiener filter," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2007, vol. 4, pp. IV–565–IV–568.

[10] 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 Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2384–2397, Dec 2015.

[11] E. Hadad, D. Marquardt, S. Doclo, and S. Gannot, "Theoretical analysis of binaural transfer function MVDR beamformers with interference cue preservation constraints," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2449–2464, Dec 2015.

[12] T. Lotter and P. Vary, "Dual-channel speech enhancement by superdirective beamforming," *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. 175–175, 2006.

[13] M. Azarpour and G. Enzner, "Binaural noise reduction via cue-preserving MMSE filter and adaptive-blocking-based noise PSD estimation," *EURASIP Journal on Advances in Signal Processing*, vol. 2017, no. 1, pp. 49, Jul 2017.

[14] M. Zohourian, G. Enzner, and R. Martin, "Binaural speaker localization integrated into an adaptive beamformer for hearing aids," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 515–528, March 2018.

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

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

[17] 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 Journal on Advances in Signal Processing*, vol. 2009, pp. 6, 2009.

[18] M. Souden, J. Chen, J. Benesty, and S. Affes, "Gaussian model-based multichannel speech presence probability," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 1072–1077, July 2010.

[19] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, April 2017.

[20] P. Vary and R. Martin, *Digital speech transmission: Enhancement, coding and error concealment*, John Wiley & Sons, 2006.

[21] R. Martin and G. Enzner, "Speech enhancement in hearing aids - from noise suppression to rendering of auditory scenes," in *Proc. IEEE Convention of Electrical and Electronics Engineers in Israel*, 2008, pp. 363–367.

[22] C.P. Brown and R.O. Duda, "A structural model for binaural sound synthesis," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 5, pp. 476–488, Sep 1998.

[23] D. Malah, R. V Cox, and A. J. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*. IEEE, 1999, vol. 2, pp. 789–792.

[24] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[25] P. Kabal, "TSP speech database," *McGill University, Database Version*, vol. 1, no. 0, pp. 09–02, 2002.

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

[27] C.H. Taal, R.C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, Sept 2011.

[28] C. Févotte, R.I. Gribonval, E. Vincent, et al., "BSS_EVAL toolbox user guide–revision 2.0," 2005.