ON THE LIMITATIONS OF BINAURAL REPRODUCTION OF MONAURAL BLIND SOURCE SEPARATION OUTPUT SIGNALS

Klaus Reindl, Walter Kellermann

Multimedia Comm. and Signal Proc.
University of Erlangen-Nuremberg
Cauerstr. 7, 91058 Erlangen, Germany
Email: {reindl, wk}@LNT.de

Mengqiu Zhang

Applied Signal Processing Group
The Australian National University
Canberra ACT 0200, Australia
Email: karan.zhang@cecs.anu.edu.au

ABSTRACT

In this contribution we analyze the binaural reproduction of an acoustic scene from monaural blind source separation (BSS) outputs for headphone applications. In practice, a perfect separation of the individual sources cannot be achieved, there are always residual components in the BSS outputs which constitute a major problem for binaural reproduction of an acoustic scene. We derive a necessary condition for the required signal-to-interference-ratio (SIR) at the BSS outputs so that a binaural reproduction of the source signals is possible up to a given tolerable error based on a simplified free-field model. The theoretical findings are verified by simulations with speech signals.

Index Terms— Source separation, binaural reproduction, HRTF models, headphone rendering

1. INTRODUCTION

The human auditory system is a binaural system with remarkable capabilities. With two ears, humans are able to localize and separate sound sources, and they are able to concentrate on a single speaker in a cocktail party, where many talkers are simultaneously active and also background noise is present. Moreover, the binaural auditory system is known to improve speech intelligibility under such conditions [1]. These phenomena demonstrate the importance of binaural hearing, and hence, the development of algorithms for binaural processing and/or rendering has attracted significant attention in recent years. Typical applications for binaural audio include computer gaming, postprocessing of live music recordings, hearing aids, or multiparty teleconferencing. Binaural audio can be rendered via loudspeakers or via headphones. In this contribution we primarily focus on a binaural headphone reproduction, where multiple monaural source separation output signals are to be rendered, in contrast to, e.g., hearing aids, where usually only a single monaural signal is to be rendered. An acoustic reproduction of blind audio source separation output signals was evaluated by extensive listening tests [2]. However, due to the binaural reproduction of multiple BSS output signals, 'aliased' sources appear because of an imperfect separation of the original source components. Therefore, we analyze the limits for a binaural reproduction of separated monaural signals within a given error from the desired directions of arrival (DOAs) of the individual sources and derive a necessary condition for the required SIR at the BSS outputs. Note that we do not consider a psychoacoustical evaluation of the virtual widening of the reconstructed source signals resulting from aliased sources.

2. SIGNAL MODEL

A general signal model for binaural reproduction of monaural output signals of BSS algorithms is depicted in Fig. 1. Lowercase boldface characters represent (column) vectors capturing signals or the filters of Multiple-Input-Single-Output (MISO) systems. Accordingly, Single-Input-Multiple-Output (SIMO) systems are described by row vectors. Matrices denoting Multiple-Input-Multiple-Output (MIMO) systems are represented by uppercase boldface characters. The superscripts $(\cdot)^*$, $\{\cdot\}^T$, and $\{\cdot\}^H$ denote complex conjugation, vector or matrix transposition, and conjugate transposition, respectively.

Due to reverberation in the acoustic environment, Q point source signals $s_q,\ q \in \{1, \ldots, Q\}$ are filtered by a MIMO mixing system modeled by finite impulse response (FIR) filters. Using the discrete-time Fourier transform (DTFT), the frequency-domain representation of the acquired microphone

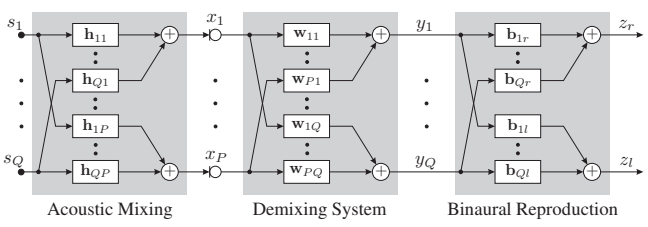


Fig. 1. Binaural reproduction of BSS output signals

signals is expressed as

$$x_p(\mathrm{e}^{\mathrm{j}\Omega}) = \sum_{q=1}^{Q} h_{qp}^*(\mathrm{e}^{\mathrm{j}\Omega}) s_q(\mathrm{e}^{\mathrm{j}\Omega}), \quad p \in \{1,\ldots,P\}, \quad (1)$$

where $\Omega = 2\pi f/f_\mathrm{s}$ is the normalized frequency, and f_s denotes the sampling frequency. The acoustic paths between the q-th source and the p-th microphone are described by frequency responses $h_{qp}(\mathrm{e}^{\mathrm{j}\Omega})$ representing FIR models with typical orders of several thousands. Conventional source separation algorithms aim at finding a corresponding demixing system so that the outputs represent estimates of the individual source signals. The output signals $y_q,\ q \in \{1,\ldots,Q\}$ are described in the DTFT domain by

$$y_q(\mathrm{e}^{\mathrm{j}\Omega}) = \sum_{p=1}^{P} w_{pq}^*(\mathrm{e}^{\mathrm{j}\Omega}) x_p(\mathrm{e}^{\mathrm{j}\Omega}), \quad q \in \{1,\ldots,Q\}, \quad (2)$$

where $w_{pq}(\mathrm{e}^{\mathrm{j}\Omega})$ denotes the current weights of the MIMO filter taps from the p-th sensor channel x_p to the q-th output channel y_q. A binaural reproduction of the monaural output signals y_q can be obtained by the binaural reproduction system described by the weights b_{qo} from the q-th output of the separation algorithm to the o-th output $o \in \{l, r\}$ of the reproduction system as follows:

$$z_o(\mathrm{e}^{\mathrm{j}\Omega}) = \sum_{q=1}^{Q} b_{qo}^*(\mathrm{e}^{\mathrm{j}\Omega}) y_q(\mathrm{e}^{\mathrm{j}\Omega}), \quad o \in \{l, r\}. \quad (3)$$

To simplify notation, the frequency-dependency $(\mathrm{e}^{\mathrm{j}\Omega})$ is omitted in the rest of the paper as long as ambiguities are precluded. Using vector/matrix notation, the acoustic mixing and the demixing, respectively, can be compactly written as

$$\mathbf{x} = \mathbf{H}^H \mathbf{s}, \quad (4)$$

$$\mathbf{y} = \mathbf{W}^H \mathbf{x} = \mathbf{W}^H \mathbf{H}^H \mathbf{s} = \mathbf{C}^H \mathbf{s}, \quad (5)$$

where the DTFT-domain signal vectors are defined as $\mathbf{s} = [s_1,\ldots,s_Q]^T$, $\mathbf{x} = [x_1,\ldots,x_P]^T$, and $\mathbf{y} = [y_1,\ldots,y_Q]^T$, respectively. The DTFT-domain mixing matrix $\mathbf{H}$ is defined as $\mathbf{H} = [\mathbf{h}_1 \ \cdots \ \mathbf{h}_P]$, where $\mathbf{h}_p,\ p \in \{1,\ldots,P\}$ is given as $\mathbf{h}_p = [h_{1p} \ \cdots \ h_{Qp}]^T$. The $P \times Q$ demixing matrix $\mathbf{W}$ and the overall $Q \times Q$ MIMO system matrix $\mathbf{C} = \mathbf{HW}$ of the separation algorithm are defined similarly to $\mathbf{H}$. A binaural reproduction of the monaural output signals $y_q,\ q = \{1,\ldots,Q\}$ of the source separation algorithm can be obtained as

$$\mathbf{z} = \mathbf{B}^H \mathbf{y} = \mathbf{B}^H \mathbf{W}^H \mathbf{H}^H \mathbf{s} = \mathbf{T}^H \mathbf{s}, \quad (6)$$

where the spatial reproduction system matrix $\mathbf{B}$ is defined as

$$\mathbf{B} = \begin{bmatrix} b_{1l} & b_{1r} \\ \vdots & \vdots \\ b_{Ql} & b_{Qr} \end{bmatrix}, \quad (7)$$

and the total system response $\mathbf{T} = \mathbf{HWB}$ is defined analogously to (7). If only one BSS output signal y_q is used for rendering (like in hearing aids), then $\mathbf{B}$ degenerates to a single row, and all other entries are equal to zero.

3. THEORETICAL LIMITS OF A BINAURAL REPRODUCTION

A detailed analysis is provided for the binaural reproduction capability of monaural BSS output signals according to the signal model introduced in Section 2. In the following, a determined case is assumed, where the number of simultaneously active sources is equal to the number of available sensors, i.e., $P = Q$.

When an acoustic scenario should be reconstructed from BSS output signals, a minimum separation performance is required to minimize aliased source components, and hence, to guarantee a binaural reproduction of a desired acoustic scene with a given or inaudible error. First of all, let us assume a perfect separation of the individual source signals by a BSS algorithm, which is described by

$$\mathbf{C} - \mathrm{diag}\{\mathbf{C}\} = \mathbf{0}, \quad (8)$$

where the operator $\mathrm{diag}\{\cdot\}$ applied to a square matrix sets all off-diagonal elements to zero. According to the condition given in (8), and assuming that the scaling and permutation ambiguities of BSS are resolved, the ideal demixing matrix results in

$$\mathbf{W}_\mathrm{ideal} = \mathrm{adj}\{\mathbf{H}\}, \quad (9)$$

where $\mathrm{adj}\{\cdot\}$ denotes the adjoint of a matrix. Consequently, the total system response $\mathbf{T}$ together with $\mathbf{B}$ and (9) results in

$$\mathbf{T}_\mathrm{ideal} = \mathbf{H}\mathbf{W}_\mathrm{ideal}\mathbf{B} = \det\{\mathbf{H}\}\mathbf{B}, \quad (10)$$

with $\det\{\cdot\}$ representing the determinant of a square matrix. Due to a perfect separation aliasing sources cannot appear because of the diagonal structure of the overall system matrix $\mathbf{C}$. As the weighting given by $\det\{\mathbf{H}\}$ in (10) is the same for both outputs, it may only cause a coloration of the binaural output signals but no distortion of binaural cues, e.g., interaural level differences and interaural time differences, respectively, which are the main cues for localizing sources in the horizontal plane [3]. However, for real acoustic environments, a perfect separation is not possible and can only be approximated. As a result, the off-diagonal elements of the overall system matrix $\mathbf{C}$ are not equal to zero as required by (8). In this case, the overall system matrix $\mathbf{C}$ can be written as

$$\begin{array}{rcccc} \mathbf{C} & = & \mathrm{diag}\{\mathbf{C}\} & + & \mathrm{offdiag}\{\mathbf{C}\} \\ & = & \mathbf{C}_\mathrm{ideal} & + & \mathbf{A} \end{array}, \quad (11)$$

where the operator offdiag sets all diagonal elements of a square matrix to zero. Using the reconstruction system $\mathbf{B}$ together with the overall system matrix of the separation algorithm (11), the total system response $\mathbf{T}$ results in

$$\mathbf{T} = \mathbf{C}\,\mathbf{B} = \mathbf{C}_\mathrm{ideal}\,\mathbf{B} + \mathbf{A}\,\mathbf{B} = \mathbf{T}_\mathrm{ideal} + \mathbf{T}_\mathrm{res}. \quad (12)$$

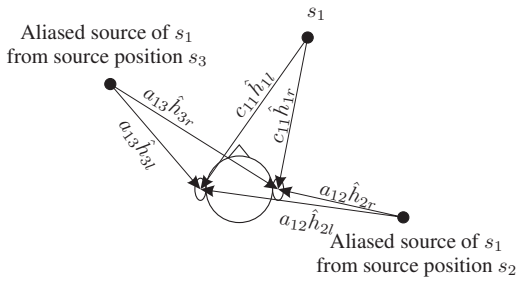


Fig. 2. Reconstruction of a monaural signal s_1 obtained from a BSS algorithm for an arbitrary scenario with three sources

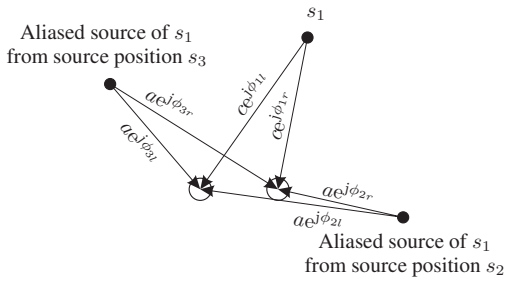


Fig. 3. Free-field model for the reconstruction of a monaural signal s_1 obtained from a BSS algorithm for an arbitrary scenario with three sources

Considering, e.g., the binaural reproduction of source s_1 for a desired acoustic scene with three sources as shown in Fig. 2, due to a non-optimum separation, the signal components s_1 will not only be reconstructed at a desired direction, but also at all remaining $Q-1$ directions (for the considered example: $Q-1=2$, and denoted as source positions s_2 and s_3), which represent aliased sources of s_1 and are described by $\mathbf{T}_{\mathrm{res}}$ in (12). Due to the superposition of the same (weighted) signal components arriving from the desired and aliased directions, a shift in position accompanied by a diffusion of the rendered signal is expected. The resulting deviation from the desired position (which is the main focus here) depends on the power of the aliased components of that signal arriving from the aliasing directions. In order to determine the deviation of the reproduced signal from the desired direction and to derive the necessary SIR at the BSS outputs so that this deviation remains within some given limits, the cross power spectral density (PSD) between the binaural output signals z_l and z_r is analyzed. To simplify the analysis, we assume a free-field model, i.e., the binaural reproduction system $\mathbf{B}$ is described by pure delays and head-shadowing effects are neglected. The simplified model of the problem depicted in Fig. 2 is shown in Fig. 3. Moreover, it is assumed that all sources can be separated equally well, i.e., $[\mathbf{C}]_{qq}=c$, and $[\mathbf{A}]_{ik}=a$. According to (12) and the model shown in Fig. 3, the reproduced source signal s_q is given as

$$z_o = s_q\left(c^*\mathrm{e}^{-\mathrm{j}\phi_{qo}} + a^*\sum_{\substack{i=1\\ i\neq q}}^{Q}\mathrm{e}^{-\mathrm{j}\phi_{io}}\right), \quad o\in\{l,r\}. \tag{13}$$

Taking the normalized cross PSD of the two output signals, we obtain

$$\frac{\hat{S}_{z_l z_r}}{\hat{S}_{s_q s_q}} = \left(c^*\mathrm{e}^{-\mathrm{j}\phi_{ql}} + a^*\sum_{\substack{i=1\\ i\neq q}}^{Q}\mathrm{e}^{-\mathrm{j}\phi_{il}}\right)\times \left(c\mathrm{e}^{\mathrm{j}\phi_{qr}} + a\sum_{\substack{i=1\\ i\neq q}}^{Q}\mathrm{e}^{\mathrm{j}\phi_{ir}}\right), \tag{14}$$

where $\hat{S}_{z_l z_r}$ and $\hat{S}_{s_q s_q}$ denote the cross PSD of the binaural output signals and the auto PSD of the source signal s_q, respectively. The argument of the cross PSD given in (14) will result in $\phi_q \pm \Delta\phi_q$, where ϕ_q represents the desired DOA of the reconstructed signal s_q and $\Delta\phi_q$ denotes the error according to the aliasing components. In order to achieve a reconstruction within a given error, we require $\Delta\phi_q \leq \Delta\phi_{\mathrm{tol}}$, i.e., the error needs to be lower than or equal to a tolerable angle $\Delta\phi_{\mathrm{tol}}$. From (14) and the requirement $\Delta\phi_q \leq \Delta\phi_{\mathrm{tol}}$, we can derive

$$\tan\left(\Delta\phi_{\mathrm{tol}}\right) \geq \frac{\sin(|\phi_q-\phi_m|)}{c/(a\cdot m)+\cos(\phi_q-\phi_m)}, \tag{15}$$

with

$$m = \sqrt{Q-1 \;+\; 2\sum_{\substack{i=1\\ i\neq q}}^{Q-1}\sum_{k=i+1}^{Q}\cos(\phi_i-\phi_k)}, \tag{16}$$

$$\phi_m = \arctan\left(\sum_{\substack{i=1\\ i\neq q}}^{Q-1}\sin(\phi_i)\Big/\sum_{\substack{i=1\\ i\neq q}}^{Q-1}\cos(\phi_i)\right), \tag{17}$$

where m and ϕ_m result from the interaction of the $Q-1$ aliasing signal components. The frequency-dependent SIR for output channel q of a BSS algorithm can be derived as

$$\mathrm{SIR}_q = \frac{|c|^2}{|a|^2}\frac{\hat{S}_{s_q s_q}}{\sum\limits_{\substack{i=1\\ i\neq q}}^{Q}\hat{S}_{s_i s_i}}, \quad q\in\{1,\ldots,Q\}, \tag{18}$$

assuming mutually uncorrelated source signals and that all signals are separated equally well. Solving (15) for c/a and inserting the result into (18), the required subband SIR for the BSS outputs so that a binaural reproduction is possible within a given error $\Delta\phi_{\mathrm{tol}}$ is derived as

$$\mathrm{SIR}_q \geq m^2\frac{\sin^2(|\phi_q-\phi_m|-\Delta\phi_{\mathrm{tol}})}{\sin^2(\Delta\phi_{\mathrm{tol}})}\frac{\hat{S}_{s_q s_q}}{\sum\limits_{\substack{i=1\\ i\neq q}}^{Q}\hat{S}_{s_i s_i}} \tag{19}$$

$$\approx \frac{m^2}{Q-1}\frac{\sin^2(|\phi_q-\phi_m|-\Delta\phi_{\mathrm{tol}})}{\sin^2(\Delta\phi_{\mathrm{tol}})}, \tag{20}$$

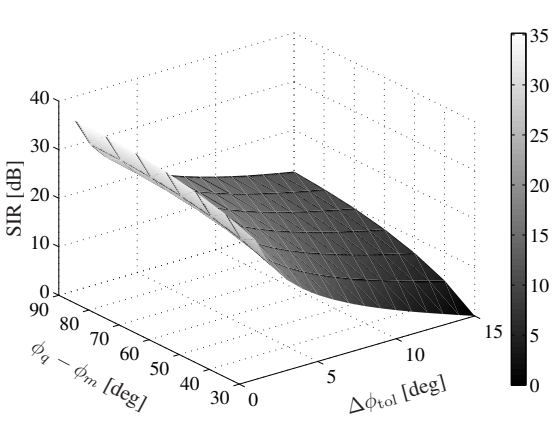
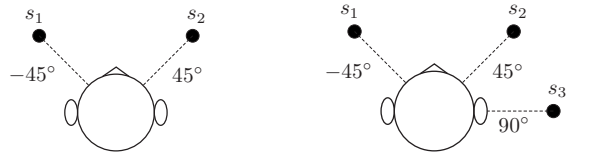


Fig. 4. Theoretical SIR at the BSS outputs required for a binaural reproduction of an arbitrary scenario with respect to a given deviation $\Delta\phi_{\text{tol}}$

where (20) is obtained assuming mutually uncorrelated white signals with equal power. The theoretical SIR (20) required at the BSS output to allow a binaural reproduction of the monaural output signals is illustrated in Fig. 4 with respect to the difference between the desired DOA ϕ_q and the resulting DOA ϕ_m according to (17), and a given error $\Delta\phi_{\text{tol}}$. It can be verified that in order to achieve a binaural reproduction of a signal s_q in an arbitrary scenario within a fixed given error $\Delta\phi_{\text{tol}}$, the required SIR at the BSS output increases with an increasing difference between the desired DOA ϕ_q and ϕ_m. Besides, for a particular scenario, i.e., $\phi_q - \phi_m$ is fixed, the required SIR at the BSS outputs can decrease for an increasing given error $\Delta\phi_{\text{tol}}$.

4. EXPERIMENTAL VERIFICATION

In order to verify the theoretical findings in Section 3, we consider in the following the binaural reproduction of different multispeaker scenarios from monaural source separation output signals. The scenarios are depicted in Fig. 5. The source signals s_1 - s_3 represent mutually uncorrelated speech signals of length 10s at a sampling rate of 16kHz. For the mixing system $\mathbf{H}$, reverberation times of $T_{60} \approx 50$ms and $T_{60} \approx 250$ms were considered. The impulse responses were measured in a low reverberation chamber and a living-room-like environment. In the following, we focus on a verification of the required SIR for monaural BSS output signals so that a desired binaural reproduction can be achieved. No real BSS scheme for source separation was applied. Instead, the monaural BSS output signals are simulated by adding the desired signals of output q and residuals to simulate different SIRs. For the different SIRs, the monaural signals are then binaurally reproduced assuming that the reproduction system $\mathbf{B}$ is given by pure delays. Analyzing the crosscorrelation at the binaural outputs $z_o, o \in \{l, r\}$, the corresponding deviation from the



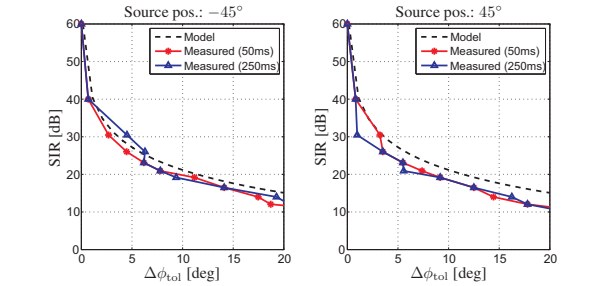
(a) Two-source scenario (b) Three-source scenario

Fig. 5. Scenarios to be reproduced from monaural source separation output signals

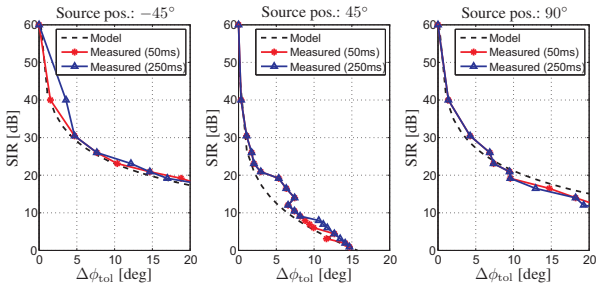
desired position was obtained. The results for the scenarios shown in Fig. 5 are illustrated in Fig. 6 together with the theoretical limit (19). They demonstrate that the model prediction matches the obtained results very well for both conditions. The fact that the measured limits for the two reverberation times are very similar is explained by the fact that throughout the analysis we did not impose constraints on the acoustic mixing $\mathbf{H}$. Only the binaural reconstruction system is constrained in terms of solely reproducing interaural delays to achieve a virtual spatial reproduction.

For a high-quality binaural reproduction of monaural signals, rather than only reconstructing interaural delays as considered above, it is important to also reconstruct level differences and monaural cues. This information is included in head-related transfer functions (HRTFs) or models thereof [4]. Therefore, we also evaluate a reproduction system $\mathbf{B}$ described by HRTF models [4]. The same experiments are performed as discussed above and compared to the theoretical limit which is, however, only based on a free-field model. The obtained results with HRTF models are illustrated in Fig. 7. It can be verified that the measurement limits have the same trend as the theoretical limits. However, due to the fact that the free-field model does not take into account scattering at the user's head, significant differences are noted. It can also be verified that due to head shadowing, the required SIR at the BSS outputs is much lower than the predicted requirement of the free-field model, i.e., the requirements for the SIR at the monaural BSS outputs are reduced compared to the free-field case, which is plausible due to the fact that head shadowing provides additional separation. From Figs. 6a and 7a it can be seen that the measured limits for the source located at $45°$ are slightly lower than for the source position at $-45°$. This non-symmetric results are due to the fact that the measured impulse responses used for the mixing system $\mathbf{H}$ are not symmetric, because of the asymmetry of the acoustic environment.

From the theoretical as well as the measurement limits, it can be seen that for a small given deviation $\Delta\phi_{\text{tol}}$, a relatively high separation performance in terms of the output SIR of BSS is required. In reality, a separation performance of up to 20dB and up to 15dB can be expected for reverberation times of 50ms and 250ms, respectively. Correspondingly, a reconstruction of the two source scenario (Fig. 5a) would only be possible with a deviation of $\Delta\phi_{\text{tol}} > 10°$ for low reverberant conditions and a deviation of $\Delta\phi_{\text{tol}} > 15°$ for moderately re-

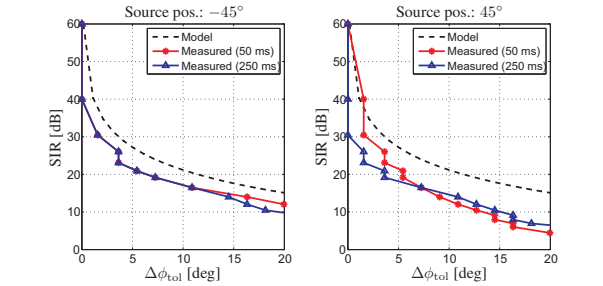


(a) Two-source scenario

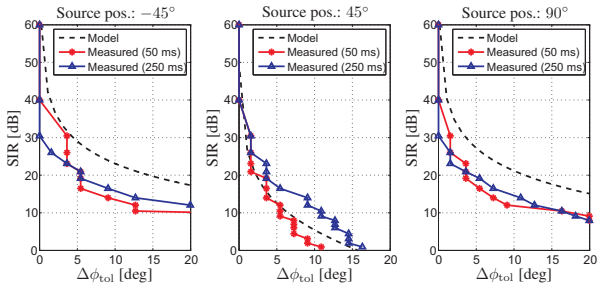


(b) Three-source scenario

Fig. 6. Comparison of the limits of a binaural reproduction of monaural BSS outputs using a free-field reconstruction system.



(a) Two-source scenario



(b) Three-source scenario

Fig. 7. Comparison of the limits of a binaural reproduction of monaural BSS outputs using a reconstruction system based on HRTF models.

verberant conditions (assuming the reconstruction system is realized by delays). However, this strongly depends on the scenario as can be verified from Fig. 6. The more realistic measurements in Fig. 7 suggest smaller $\Delta\phi_{\text{tol}}$ for these SIRs ($\Delta\phi_{\text{tol}} > 5°$ for $T_{60} \approx 50$ms and $\Delta\phi_{\text{tol}} > 10°$ for $T_{60} \approx 250$ms).

5. CONCLUSION

In this paper we analyzed the theoretical limits for a spatial reproduction of monaural BSS output signals for headphone applications, where multiple BSS output signals are rendered in order to reproduce or manipulate an acoustic scene. A major problem for binaural rendering are the residual components contained in the BSS output signals. The residuals contained in the monaural BSS output signals lead to aliasing sources after binaural reproduction of these signals so that a rendered virtual source is perceived at a different location. Allowing a given tolerable error of the desired DOA of the rendered virtual source, the residuals in the BSS outputs have to be below a certain threshold. Based on a simplified free-field model, we derived the required BSS performance in terms of the SIR at its outputs. Experiments with speech signals demonstrated the validity of the model. However, using HRTF models as binaural reproduction system instead of a free-field model, significant deviations between the theoretical model and the experimental results are observed, which is due to the fact that the derived theoretical limit does not account for the extra separation by head shadowing and scattering. The experimental results confirmed that accounting for HRTFs for a binaural reproduction of BSS output signals reduces the requirements for the separation algorithm, i.e., a lower separation performance can be tolerated to achieve a reconstruction within the same given deviation as with a free-field model. Further investigations will focus on an incorporation of an approximate spherical head model [5] to analyze for the first time the interaction of ILDs and ITDs for the requirement of the BSS performance. Moreover, for a psychoacoustical evaluation, listening test should be conducted to account for the the virtual widening of the source signals resulting from aliasing sources which are created when rendering multiple BSS output signals.

6. REFERENCES

[1] M. L. Hawley, R. Y. Litovsky, and J. F. Culling, "The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer," *J. Acoust. Soc. Am.*, vol. 115, no. 2, pp. 833–843, 2004.

[2] T. Kastner, "The influence of texture and spatial quality on the perceived quality of blindly separated audio source signals," in *129th AES Convention*, San Francisco, CA, Nov. 2010.

[3] J. Blauert, *Spatial hearing: The psychophysics of human sound localization*, MIT Press, 1997.

[4] W. Zhang, T. D. Abhayapala, R. A. Kennedy, and R. Duraiswami, "Insights into head related transfer function: Spatial dimensionality and continuous representation," *J. Acoust. Soc. Am.*, vol. 127, no. 4, pp. 2347–2357, April 2010.

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