Upmixing Stereo Signals for Multichannel Audio Reproduction

Arne Jacobsen, Julian Grosse, Steven van de Par
Acoustics Group, Carl von Ossietzky University, 26111 Oldenburg
Email: arne.jacobsen@uni-oldenburg.de

Introduction

Nowadays it is common that music pieces include a dedicated multichannel down mix in addition to the still popular stereo down mix. But there will be cases where this is not possible, e.g. loss of raw recordings or even death of artists. For this purpose there has to be the opposite for the down mix procedure, namely an upmixing of existing stereophonic audio content. In this study an upmixing algorithm that will decompose the stereo signal into a direct signal and a diffuse signal has been utilized. These decomposed signals will be rendered over a multichannel reproduction system that consists of two standard HiFi front loudspeakers and two dipole rear loudspeakers. This four channel audio reproduction system aims at an perceptually accurate reproduction of recorded sound fields, whereas a direct signal will be rendered over the front loudspeakers and a reverberant signal over the rear loudspeakers. In addition to this there is a perceptually motivated optimization method that compensates perceptively relevant aspects of the reproduction room.

Method

As depicted in Figure 1 the approach of this study can be separated into three parts. The first part decomposes the stereo signal into a direct and diffuse signal. The two subsequent parts are optimizing the signals with respect to the reproduction room.

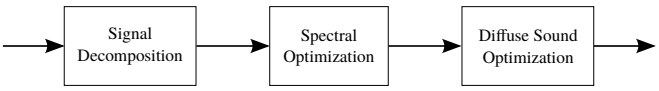

Figure 1: General approach for the proposed method. Signal decomposition into a direct and a diffuse part, spectral optimization with respect to the reproduction room and diffuse sound optimization.

Signal Decomposition

The signal decomposition, which has been done according to [1], generates a *direct* signal s and a *diffuse* signal d from a stereophonic input. The underlying signal model can be expressed in simplified form as follows:

$$\begin{aligned} x_L &= \sin(\gamma)s + d \\ x_R &= \cos(\gamma)s - d \end{aligned} \tag{1}$$

Where the subscripts L and R are representing the left and right stereo channel. In order to calculate the direct and diffuse signal the angle γ needs to be determined. For this the DFT bins are grouped to non-overlapping bands in order to approximate an ERB-scale resolution. After this the angle γ of the phantom source will be estimated for each band. Therefore the normalized cross-correlation coefficient will be determined by

$$\rho[b] = \Re \left(\frac{\sum_{k=m}^{B} X_L[k] X_R^*[k]}{\sigma_{X_L}[b]\sigma_{X_R}[b]} \right) \quad , \tag{2}$$

with

$$\sigma_X[b] = \sqrt{\sum_{k=m}^{B} X[k]X^*[k]} \quad . \tag{3}$$

In this case $\Re(z)$ denotes the real part of z and z^* the complex conjugate of z. The lower limit m of the sum in (2) and (3) represents the DFT bin that corresponds to the lower limit of the specific auditory band b whereas B marks the upper boundary. Based on some goniometric functions (see Eq. (6), (7) and (9) in [1]) the normalized cross-correlation coefficient will be transformed to the virtual source angle γ. After this the direct signal and diffuse signal can be calculated as follows:

$$\begin{aligned} S[k,n] &= \frac{X_L[k,n] + X_r[k,n]}{\sin(\gamma[k,n]) + \cos(\gamma[k,n])} \\ D[k,n] &= X_L[k,n] - \sin(\gamma[b,n])S[k,n] \quad . \end{aligned} \tag{4}$$

In this case the capital letters are denoting the time-frequency representation of the corresponding signals, whereas k represents the DFT bins and n the time frames. By means of a small additional step (see (5)) the direct signal S can be transformed into a two-channel signal that will be send to the front loudspeakers.

$$\begin{aligned} S_L[k,n] &= S[k,n]\sin(\gamma[b,n]) \\ S_R[k,n] &= S[k,n]\cos(\gamma[b,n]) \end{aligned} \tag{5}$$

A schematic diagram for the described processing steps and signal representations can be seen in Figure 2.

Spectral Optimization

For the spectral optimization an approach has been used that was initially introduced in [3]. The aim is to achieve a flat ATF, which is an acronym for auditory transfer function that has been introduced in [3]. This measure indicates the amount of energy in each auditory band. In order to apply this to the optimization of the reproduction room the room impulse response for the four utilized loudspeakers needs to be measured. With this the

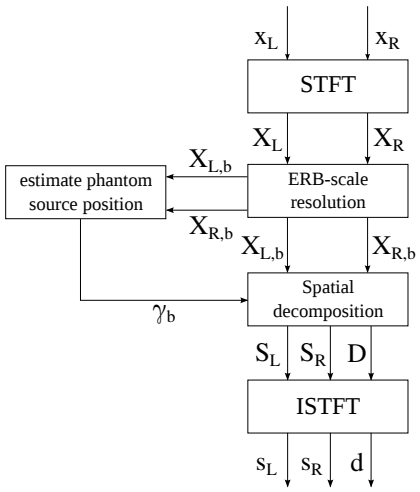

Figure 2: Signal decomposition from stereo into direct and diffuse signal by estimating a dominant source in each auditory channel based on energetic ratios

spectral optimization procedure can be mathematically expressed as follows:

$$\mathrm{ATF}(H_{LS}[k]) \approx \mathrm{ATF}(H_{ref}[k]) \quad . \tag{6}$$

The method for separating the frequency range into auditory bands is identical to the one that has been used for the signal decomposition. Thus an ATF is calculated according to (7).

$$\mathrm{ATF}(H_{LS}[k]) = \frac{1}{K} \sum_{k=m}^{B} H_{LS}[k] H_{LS}^{*}[k] \tag{7}$$

Where K represents the number of DFT bins that build the corresponding band. For achieving an ATF that will yield the reference, a scaling factor g needs to be introduced, i.e. from (6) it follows:

$$g[b] = \sqrt{\frac{\mathrm{ATF}(H_{ref}[k])}{\mathrm{ATF}(H_{LS}[k])}} \quad . \tag{8}$$

With this a filter for each loudspeaker can be designed that will adjust the corresponding ATF.

Diffuse Sound Optimization

The diffuse sound optimization is motivated as well by [3], whereas the aim is to achieve a certain interaural cross-correlation (IACC) for the reverberant sound field. This procedure is as well based on the assumption that the human auditory system includes a process of auditory filtering. Thus we introduce the auditory cross-correlation (ACC), which describes the cross-correlation for every center frequency of the auditory bands. In this approach the aim is to achieve an ACC of the dipole signals that is equal or at least close to a predefined reference. This reference is given by the interaural cross-correlation coefficient (IACC) that can be measured with two microphones (spaced at distance r) in an ideal diffuse sound field, which can also be expressed mathematically as follows:

$$\mathrm{IACC}_{ref} = \frac{\sin(kr)}{kr} \quad . \tag{9}$$

In this case k refers to the wave number. By adopting this to the proposed ACC concept (9) turns into:

$$\mathrm{ACC}_{ref} = \frac{\sin(k_{b_c} r)}{k_{b_c} r} \quad . \tag{10}$$

Where k_{b_c} denotes the wave number that corresponds to the specific auditory band. Due to the fact that the decomposition algorithm only delivers a single diffuse signal a decorrelation stage will generate two nearly uncorrelated diffuse signals $d_{dec,1}$ and $d_{dec,2}$. Thus each rear dipole loudspeaker will have a dedicated diffuse signal. Afterwards a crossmixing according to (11) will generate signals d_L and d_R which will have a cross-correlation ρ that is close to α. The case that $\rho = \alpha$ will only hold for uncorrelated signals.

$$\begin{aligned} d_L[n] &= d_{dec,1}[n] + \alpha \cdot d_{dec,2}[n] \\ d_R[n] &= d_{dec,2}[n] + \alpha \cdot d_{dec,1}[n] \end{aligned} \tag{11}$$

From here starts an iterative process which varies the value for α between -1 and 1. Once the crossmixing has been done a convolution of the diffuse signals with room impulse responses (see Figure 3) will deliver the reproduced diffuse signals. Therefore in the last step the ACC which represents the cross-correlation coefficient per auditory band can be calculated. Since this approach generates a matrix that consists of ACCs for each value of alpha an additional step applies band dependent crossmixing by using the value of α that results in an ACC that is as close as possible to the reference.

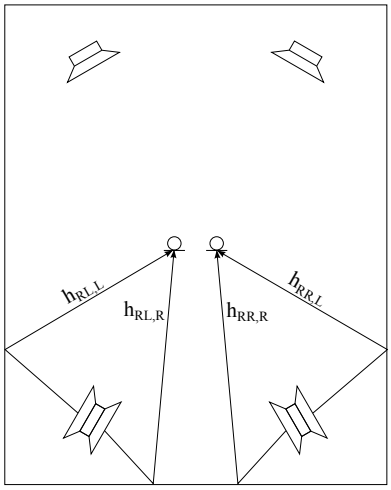

Figure 3: Setup in Which room impulse responses (RIRs) of rear dipole loudspeakers measured with two microphones mounted on a stereo bar (distance $r = 16$ cm)

Evaluation

For the subjective evaluation rating tests on a scale from 0 to 100 have been conducted by eight male subjects with normal hearing. In addition to a general *preference* test there have been queried three attributes from the Spatial Audio Quality Inventory (SAQI [4]), namely *width, presence and ghost source*. The rating has been done via a *MUSHRA-like* user interface. In general there have

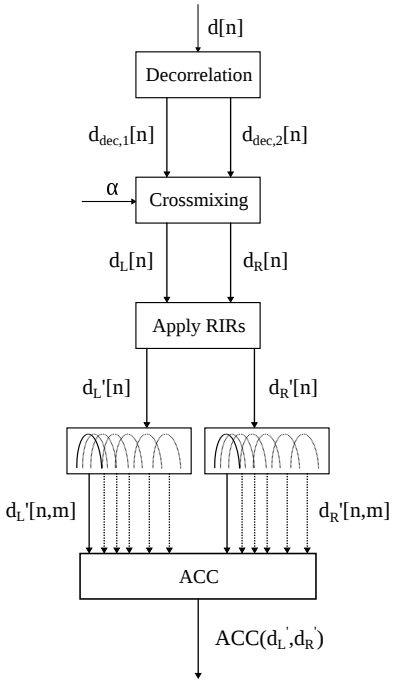

Figure 4: Diffuse sound optimization - *decorrelation* for generating two diffuse signals, *crossmixing* for achieving certain correlation, *convolution* with RIRs to investigate influences of reproduction room and *ACC* for analyzing cross-correlation in each auditory band

been ten audio excerpts from five different genres (classic, pop, rock, singer/songwriter and electro). Each attribute has been queried by five conditions (see Table 1). The results for the preference rating (see Figure 5)

Condition	Description
decomp surround	$s_{L/R}$ via front loudspeakers, $\pm d$ via standard surround rear loudspeakers
decomp dipole	$s_{L/R}$ via front loudspeakers, $\pm d$ w/ ATF (anechoic) optimization via dipoles
optimized I	$s_{L/R}$ w/ ATF optimization via front loudspeakers, $\pm d$ w/ ATF optimization via dipoles
optimized II	$s_{L/R}$ w/ ATF optimization via front loudspeakers, $d_{L/R}$ w/ ATF and ACC optimization via dipoles
stereo	$x_{L/R}$ via front loudspeakers

Table 1: Tested conditions for preference and SAQI attributes

are showing that only the *decomp dipole* condition is at least equally preferred as the *stereo* condition. The other three conditions are more or less equally rated. For the attributes *width and presence* it can be seen that again *decomp dipole* is more or less rated at the same level as *stereo*, whereas the other conditions are lower rated. The absolute ratings for the attribute ghost source (see Figure 8) have to be treated carefully because there have

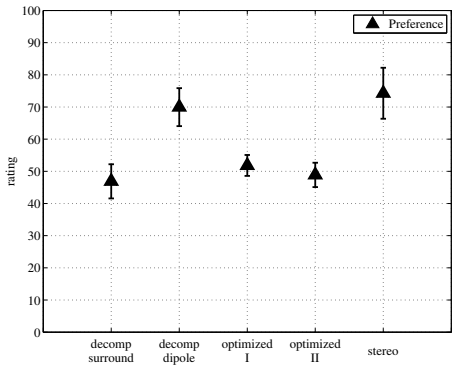

Figure 5: Results preference test - triangles representing mean across ten audio excerpts and eight test subjects, error bars indicating standard error related to number of subjects

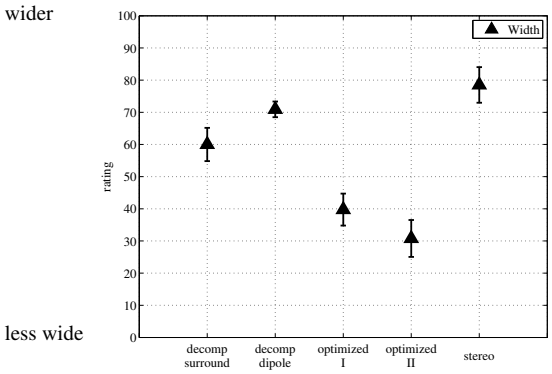

Figure 6: Results attribute width (*less wide - wider*) - triangles representing mean across three audio excerpts and eight test subjects, error bars indicating standard error related to number of subjects

been subjects that detected ghost sources at the *stereo* condition which can technically not be the case.

Discussion

Looking at the results for the attributes *preference, width and presence* there can always be seen two groups, whereas *decomp dipole* and *stereo* build the first well rated group. The other three conditions *decomp surround, optimized I and II* form the second less rated group. By the first group it can be said that reaching the same rating as the original *stereo* signal is a good result. The second group shows that neither of its condition will give a remarkable improvement compared to the *stereo* condition. The fact that the optimization of the diffuse sound field delivers no audible improvement *(optimized II)* is in line with the results of [2]. But in addition to that they have found that the spectral optimization *(optimized I)* will result in a perceivable improvement. This leads to the fact that the spectral optimization is highly dependent on the audio content, i.e. this procedure works less well on down mixed and reworked signals.

Conclusion

In this study it has been shown that the appropriated decomposition algorithm applied to stereo signals and ren-

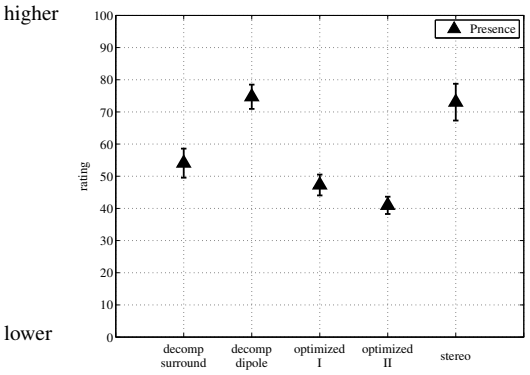

Figure 7: Results attribute presence (*lower - higher*) - triangles representing mean across three audio excerpts and eight test subjects, error bars indicating standard error related to number of subjects

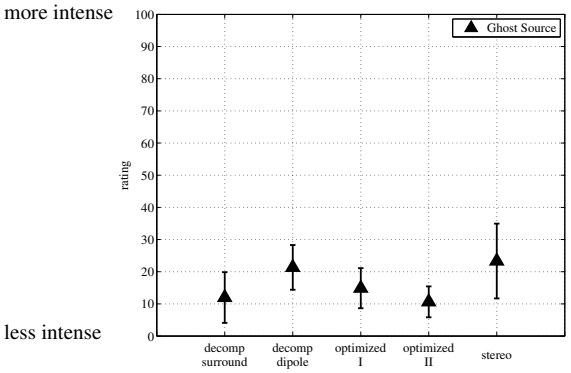

Figure 8: Results attribute ghost source (*less intense - more intense)* - triangles representing mean across three audio excerpts and eight test subjects, error bars indicating standard error related to number of subjects

dered in a multichannel reproduction environment is at least equally preferred as a classical stereo reproduction. In addition it showed that dipoles as rear loudspeakers are preferred over standard surround loudspeakers, which might leads to the assumption that dipoles can be utilized as surround loudspeakers in other multichannel audio reproduction systems, e.g. 5.1 or 7.1, as well. Furthermore there are some conceivable improvements to this work. One could be a redefining of the reference ATF in the spectral optimization in order to avoid an unnatural reproduction scenario due to large amplification at high frequencies. Another improvement could be additional front dipoles to each front loudspeaker. This will theoretically compensate the decreasing apparent source width because the frontal loudspeakers will still reproduce the stereo signal except the fact that the diffuse signal will be orthogonally reproduced to the direct signal. A final improvement could be a higher level of the rear dipole loudspeakers in order to increase the effect of presence.

References

[1] Jeroen Breebaart and Erik Schuijers. “Phantom materialization: A novel method to enhance stereo audio reproduction on headphones”. In: *Audio, Speech, and Language Processing, IEEE Transactions on* 16.8 (2008), pp. 1503–1511.

[2] Julian Grosse and Steven van de Par. “Evaluation of a perceptually optimized room-in-room reproduction method for playback room compensation”. In: (2015).

[3] Julian Grosse and Steven van de Par. “Perceptually accurate reproduction of recorded sound fields in a reverberant room using spatially distributed loudspeakers”. In: *Selected Topics in Signal Processing, IEEE Journal of* 9.5 (2015), pp. 867–880.

[4] Alexander Lindau et al. “A spatial audio quality inventory (SAQI)”. In: *Acta Acustica united with Acustica* 100.5 (2014), pp. 984–994.