# STEREO SIGNAL SEPARATION AND UPMIXING BY MID-SIDE DECOMPOSITION IN THE FREQUENCY-DOMAIN

*Sebastian Kraft, Udo Zölzer*

Department of Signal Processing and Communications
Helmut-Schmidt-University
Hamburg, Germany
sebastian.kraft@hsu-hh.de

**ABSTRACT**

An algorithm to estimate the perceived azimuth directions in a stereo signal is derived from a typical signal model. These estimated directions can then be used to separate direct and ambient signal components and to remix the original stereo track. The processing is based on the idea of a bandwise mid-side decomposition in the frequency-domain which allows an intuitive and easy to understand mathematical derivation. An implementation as a stereo to five channel upmix is able to deliver a high quality surround experience at low computational costs and demonstrates the practical applicability of the presented approach.

## 1. INTRODUCTION

The classical stereo format with a left and right speaker was developed by Blumlein in the 1930s. Although it allows to create a certain spaciousness by placing virtual sound sources on the azimuth angle between both loudspeakers, the result is still far away from a really realistic rendering of sound scenes. The main disadvantages of stereo are:

1. Impossibility to produce a real envelopment of the listener as rear sound sources are missing. These would be in particular important for a reproduction of ambient reflections to get a realistic impression of the room properties.
2. Phantom source positions are only properly reproduced for listeners in the sweet spot.

It has long been known that both problems can be solved by adding speakers in the rear for playback of ambient reflections and more speakers in the front for a stabilisation of the phantom source image. Today, surround sound is widely established in film and dedicated distribution media like DVD or Blu-Ray generally offer at least a 5.1 multi-channel sound track. In contrast, the majority of popular music is still exclusively produced and distributed in stereo and would not benefit from playback on more than two speakers. Even if more music productions would become available in multi-channel formats in the next years, there is still a huge stock of old stereo recordings.

Several approaches were developed in the past to benefit from additional loudspeakers while playing back stereo source material. The most simple ones use time-domain mixing matrices [1] together with phase shifting to generate the additional channels. More advanced upmixing algorithms usually follow a spatial source separation approach as depicted in Fig. 1. They try to split the stereo signal into a direct part and a diffuse and more or less uncorrelated ambient part. The direct signal will be repanned on the target speaker layout while the ambient signal will be equally redistributed to all speakers creating a uniform and diffuse sound field. In order to retain the perceived positions of the direct signal sources in the upmix, it is necessary to estimate their azimuth directions from the stereo signal and to incorporate this knowledge in the calculation of the repanning coefficients.

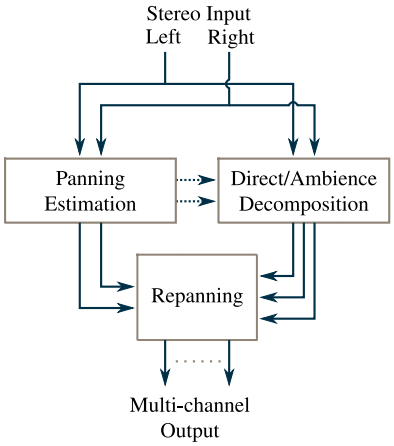



Figure 1: *Exemplary processing steps in a typical stereo to surround upmix algorithm.*

Dolby Pro Logic II [2] became quite popular in the 1990s and was intended to extract five channels from a stereo track. Basically it is a matrix encoder/decoder system but is also capable to extract decent multi-channel signals from any arbitrary stereo mix. For a simple and cost effective realization it only requires time domain operations like subtraction and addition of the left and right channels with additional phase shifts and VCAs (voltage controlled amplifiers) for simple directional steering.

Recent algorithms make use of frequency-domain processing to analyse the signal in discrete frequency bands. Avendano and Jot [3] calculate a bandwise inter-channel short-time coherence from the cross- and autocorrelations between the stereo channels which is then the basis for the estimation of a panning and ambience index [4]. Faller [5] uses a least squares method to derive an algorithm where the error between the extracted signals and a stereo signal model is minimised. Goodwin [6] examines the left and right stereo signals in a 2D vector space and extracts the direct and ambient sound with a principal component analysis. A similar geometric decomposition is described by Vickers [7] for the purpose of center extraction. An enhanced time-domain rever-

beration extraction upmixer was presented by Usher [8] where a normalised least mean squares (NLMS) algorithm is used to find a filter for the extraction of uncorrelated components. All the above algorithms [3–8] and also the one presented in this paper share a comparable stereo signal model with similar assumptions about the individual signal components.

The focus of this contribution is on a simplified mathematical description and derivation, which finally leads to a very fast and efficient implementation. It is shown that most of the processing principles, which are also known from other approaches, can be interpreted as a simple generalised mid-side decomposition which is performed in sub-bands.

In the following section 2, a typical mathematical model to describe stereo signals is derived and its connections to mid-side decomposition and principal component analysis are highlighted. Section 3 describes the extraction of the direct and ambient signals as well as the estimation of the direct signal directions. The upmixing of the separated components to a generic surround sound setup and other applications are outlined in section 4 before the results of an exemplary 2-to-5 upmix implementation are discussed in section 5. Section 6 will conclude the paper.

## 2. STEREO SIGNAL MODEL

The left and right channels of a stereo signal

$$x_L(n) = \left[\sum_{j=1}^{J} a_{L_j} \cdot d_j(n)\right] + n_L(n) \tag{1}$$

$$x_R(n) = \left[\sum_{j=1}^{J} a_{R_j} \cdot d_j(n)\right] + n_R(n) \tag{2}$$

are usually described as a weighted sum of $J$ source signals $d_j(n)$ and additive uncorrelated ambient signals $n_L(n)$ and $n_R(n)$ in the left and right channel, respectively. The weightings $a_{L_j}$ and $a_{R_j}$ of the individual sources are called panning coefficients and are bound between zero and one. Their squared sum should be equal to one ($a_{L_j}^2 + a_{R_j}^2 = 1$) to achieve a constant power and loudness independent of the actual source position. As the panning coefficients are real-valued, this model only covers intensity stereophony where the weighted sources in the left and right channels are in phase.

The time-domain signal model can directly be transformed into the frequency-domain by a short time fourier transform (STFT)

$$X_L(b,k) = \left[\sum_{j=1}^{J} a_{L_j} \cdot D_j(b,k)\right] + N_L(b,k) \tag{3}$$

$$X_R(b,k) = \left[\sum_{j=1}^{J} a_{R_j} \cdot D_j(b,k)\right] + N_R(b,k) \tag{4}$$

where $b$ and $k$ denote the block and frequency indices. Based on the two stereo channels as input, it is impossible to mathematically retrieve the sources, panning coefficients and ambient signals as the equation systems (1)-(2), (3)-(4) are highly underdetermined. However, for a sufficiently high time and frequency resolution it is a common assumption [9] that at a certain time instant $b$ and in a frequency band $k$ only a single dominant source $D_u$ is active and the contribution of other sources is close to zero

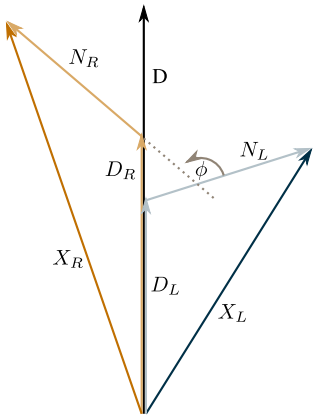



Figure 2: *The stereo signal model from (6)-(7) visualised as a complex vector diagram for a single frequency band $k$.*

($\sum_{\forall j \neq u} |D_j(b,k)| \approx 0$) . This allows to summarise the time-frequency representations of the individual sources

$$a_L(b,k) \cdot D(b,k) = \sum_{j=1}^{J} a_{L_j} \cdot D_j(b,k) \approx a_{L_u} \cdot D_u(b,k) \tag{5}$$

into a single source $D(b,k)$ and the panning coefficients to be written as a matrix $a_{L/R}(b,k)$. Moreover, the left and right ambient signal in one band $k$ can be expected to have a similar magnitude but due to different paths and reflections in the room, they likely have a different phase and can be replaced by:

$$N_L(b,k) = N(b,k), \qquad N_R(b,k) = e^{j\phi} \cdot N(b,k).$$

Combining both steps leads to a simplified signal model

$$X_L(b,k) = a_L(b,k) \cdot D(b,k) + N(b,k) \tag{6}$$

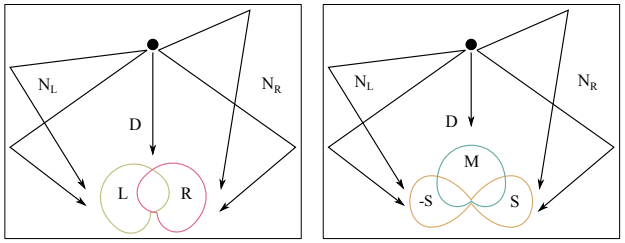
$$X_R(b,k) = a_R(b,k) \cdot D(b,k) + e^{j\phi} \cdot N(b,k) \tag{7}$$

with a highly reduced number of unknowns. For an improved readability the indices $(b,k)$ are omitted in the next sections.

Overall, the simplifications in the signal model may seem to be quite drastic and one can doubt its validity in particular for complex and high density musical recordings. But the intention of the presented model is not to allow an exact extraction and reproduction of the original source signals. The idea is to have a generic signal model that is able to describe the relation between directional and diffuse signal components for an arbitrary stereo signal. The resulting ambient and direct signal components should be free of artefacts and sound realistic, but for that purpose they do not necessarily have to be identical to the real signals before the downmix.

### 2.1. Interpretation as complex vector diagram

The signals from the model (6)-(7) are complex-valued and can be depicted in form of a vector diagram to visualise their relation in phase and magnitude. In Fig. 2 the weighted direct signal components $D_{L/R} = a_{L/R} \cdot D$ are in phase with the direct signal $D$ as the coefficients $a_{L/R}$ are only real-valued and do not alter the phase. In contrast the ambient components $N_{L/R}$ are out of phase




(a) *Left-right (XY) sterophony* (b) *Mid-side (MS) stereophony*

Figure 3: *Comparison between left-right and mid-side stereo recording techniques regarding the capturing of direct and ambient components.*

with the direct signal. Furthermore, $N_L$ and $N_R$ have a similar amplitude but a phase difference of $\phi$.

For an angle $\phi = \pi$ the vector diagram transforms into the signal model from [6] where the direct signal is estimated as the principal component and the ambient signals are orthogonal to it.

### 2.2. Interpretation as mid/side stereophony

A stereo signal is typically recorded as a left and right channel ($X_L$/$X_R$) but it can also be represented by a mid and side channel ($X_M$/$X_S$). Both variants can be converted into each other

$$X_M = 0.5 \cdot (X_L + X_R) \tag{8}$$
$$X_S = 0.5 \cdot (X_L - X_R) \tag{9}$$

$$X_L = X_M + X_S \tag{10}$$
$$X_R = X_M - X_S \tag{11}$$

by sum and difference calculations. The connection between the above stereo representations and the signal model (6)-(7) becomes apparent in Fig. 3 where two classical stereo recording setups are depicted in a simple room model.

One way to record a left-right stereo signal is to use two cardoid microphones in a XY configuration as shown in Fig. 3 a). The direct sound $D$ emitted from a central sound source will be recorded with the same intensity and phase by both microphones. The ambient signals $N_L$ and $N_R$ reach the left and right microphone with different phase but similar amplitude. This corresponds to the signal model from (6)-(7) in the case $a_L \equiv a_R$.

To record a mid-side stereo signal, a figure of eight microphone is faced sideways to capture the side signal and a cardoid micropone captures the mid signal. Placed in a room with a single central sound source as shown in Fig. 3 b), the mid microphone will nearly exclusively capture the direct signal, whereas the figure of eight mostly captures ambient reflections. Therefore, a mid-side stereo signal already implies a certain degree of separation between ambient and direct components. Having a left-right stereo signal pair as input, one can achieve a simple direct-ambience decomposition by calculating the mid and side signals with (8)-(9). Indeed, Dolby Pro Logic I and II [2,10], for example, already made use of this idea and basically obtained the center and ambient signals by sum and difference calculations of the left and right stereo channels. However, due to pure time-domain processing, their capability to separate multiple sources at the same time was limited.

## 3. SIGNAL EXTRACTION

### 3.1. Panning coefficient estimation

The panning coefficients in (6)-(7) are real-valued which means that they only create amplitude panning and do not introduce a phase difference between $X_L$ and $X_R$. Any phase shift between the left and right channel would be solely caused by an additive ambient signal with a magnitude $|N| > 0$ (which is also apparent from Fig. 2). However, for typical music mixes the amplitude of the ambient signal $N$ is far less than the amplitude of the direct signal $D$. This also means that the left and right channels

$$|X_L| \approx a_L \cdot |D| \tag{12}$$
$$|X_R| \approx a_R \cdot |D| \tag{13}$$

are sufficiently approximated by the weighted direct signal magnitude and the phase can be neglected in this relation. Rearranging and solving equations (12)-(13) with the constraint $a_L^2 + a_R^2 = 1$, the panning coefficients

$$\hat{a}_L = \frac{|X_L|}{\sqrt{|X_L|^2 + |X_R|^2}} \tag{14}$$
$$\hat{a}_R = \frac{|X_R|}{\sqrt{|X_L|^2 + |X_R|^2}} \tag{15}$$

can be estimated from $X_L$ and $X_R$.

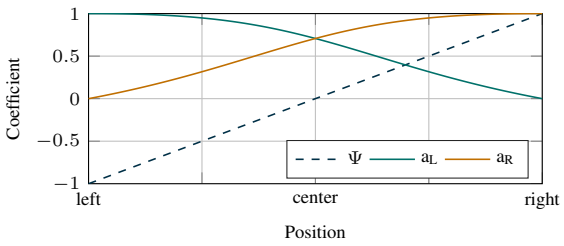
The "stereophonic law of sines" [11]

$$\frac{a_L - a_R}{a_L + a_R} = \frac{\sin\theta}{\sin {}^{\theta_0}/_2} = -\Psi \tag{16}$$

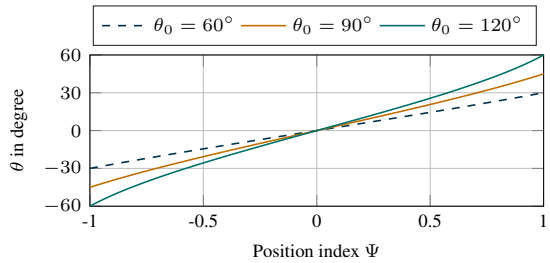
describes the perceived angle $\theta$ of a source if its amplitude is weighted by $a_{L/R}$ for playback on a left and right loudspeaker while the angle between the both speakers is defined by $\theta_0$. The normalised position index $\Psi$, ranging from $-1$ for left and $+1$ for right positions, combines the coefficients $a_{L/R}$ in a single value (Fig. 4 a)). From (12)-(13) and (16) one can derive estimates for the position index and angle

$$\hat{\Psi} = \frac{|X_R| - |X_L|}{|X_L| + |X_R|} \tag{17}$$
$$\hat{\theta} = \arcsin\left(\sin {}^{\theta_0}/_2 \cdot \frac{|X_R| - |X_L|}{|X_L| + |X_R|}\right) \tag{18}$$

based on the magnitudes of the left and right stereo channel. When plotting the perceived source angle for different values of $\theta_0$ and over the normalised position index $\Psi$ in Fig 4 b), one can see that there is a nearly linear relation for the typical stereo setup with $\theta_0 = 60°$. Hence, the normalised position $\Psi$ nicely matches the human perception and can be directly used as a linear pan-pot like control parameter to describe source positions. Comparing (17) with the mid-side decomposition mentioned in section 2.2, it appears that the position $\hat{\Psi}$ is the ratio between the side and mid component of an approximately coherent stereo signal.




(a) *Panning coefficients $a_{L/R}$ and position index $\Psi$*




(b) *Mapping between position index $\Psi$ and the perceived angle $\theta$ in dependency of the angle between two speakers $\theta_0$*

Figure 4: *Mapping between panning coefficients, position index and perceived angles.*

### 3.2. Direct and ambient signal separation

The estimated panning coefficients from the previous section can be used to solve the signal model (6)-(7) for

$$\hat{D} = \frac{X_L\, e^{j\phi} - X_R}{\hat{a}_L\, e^{j\phi} - \hat{a}_R} \tag{19}$$

$$\hat{N} = \frac{\hat{a}_L\, X_R - \hat{a}_R\, X_L}{\hat{a}_L\, e^{j\phi} - \hat{a}_R} \tag{20}$$

$$\hat{N}_L = \hat{N} \qquad = X_L - \hat{a}_L \cdot \hat{D}$$

$$\hat{N}_R = e^{j\phi} \cdot \hat{N} \qquad = X_R - \hat{a}_R \cdot \hat{D}$$

to get an estimate of the direct and ambient signal components. Currently no method is known to guess the angle $\phi$ and therefore, at the moment it is kept as an adjustable input parameter to influence the signal separation process. By setting $\phi = \pi$ the resulting left and right ambient signals are out of phase. While this may help to increase the perceived spatial depth of the ambient signals, it also causes unpleasant phase cancellations. Choosing $\phi$ in a range $\phi \in [0.5\,\pi, \pi]$ leads to less cancellations and for $\phi = 0.5\,\pi$ a maximum decorrelation between both ambient signals can be achieved. For the application of upmixing with a setup as described in the next section, an angle $\phi = 0.6\,\pi$ yielded the best balance between spatial depth and out-of-phase artefacts.

The formulas (19)-(20) already show a certain similarity with the mid-side decomposition previously described in section 2.2. Indeed, with $\hat{a}_L = \hat{a}_R = 1$ and a phase $\phi = \pi$ they become

$$\hat{D} = \frac{X_L\, e^{j\pi} - X_R}{e^{j\pi} - 1} = \frac{X_L + X_R}{2} \tag{21}$$

$$\hat{N} = \frac{X_R - X_L}{e^{j\pi} - 1} = \frac{X_L - X_R}{2} \tag{22}$$

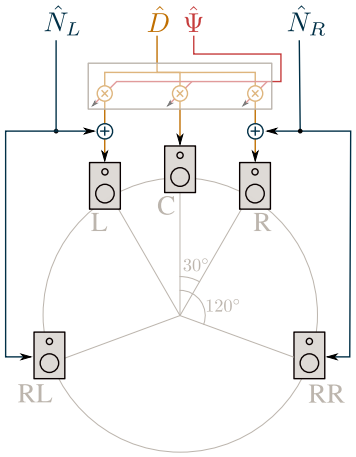



Figure 5: *Upmixing of the extracted ambient and direct signals to a five speaker surround setup.*

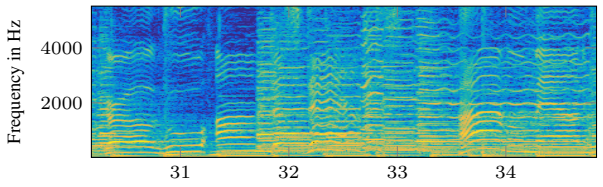
which is identical to (8)-(9) and shows that the derived direct and ambience separation essentially is a generalised mid-side decomposition. In contrast to a pure time-domain mid-side conversion, the panning coefficients in each sub-band are incorporated to allow a proper separation of direct signals which were not panned to the center.

## 4. REMIX

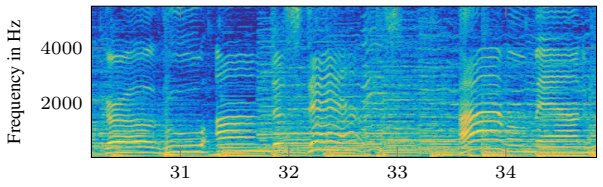
Using the separated signal components $\hat{D}, \hat{N}_L, \hat{N}_R$ and with the knowledge of the panning coefficients $\hat{a}_L, \hat{a}_R$ (or source positions $\hat{\Psi}$ and angles $\hat{\theta}$) from the previous section, it is possible to remix the original stereo signal. In the context of upmixing this would mean to redistribute the signals to a different loudspeaker arrangement. In most cases it is desired to retain the perceived source positions as they were placed in the stereo mix. However, it is also possible to widen or narrow the stereo panorama or to completely modify individual source positions. The ambient signal is equally distributed to all loudspeakers to create a diffuse sound field while the balance between ambient and direct signal can be modified.

The signal flow for a typical upmix scenario to five speakers is depicted in Fig. 5. The direct signal is played back by the three front speakers (L, C, R), whereas the corresponding weights for each front channel are obtained by *Vector Base Amplitude Panning* (VBAP) [12]. The left and right ambient signals are added to the four corner speakers (RL, RR, L, R). For a highly diffuse ambient sound field it is necessary to establish a low correlation between all ambient loudspeaker signals as pointed out by Kendall [13]. While the left and right ambient signals already have a low correlation if the angle $\phi$ is selected properly (cf. section 3.2), only the front and rear ambient signals on each side are fully correlated. In the most simple case these could be decorrelated by adding a small delay to the rear channels. More complex approaches in the frequency-domain are for example proposed by [13, 14].

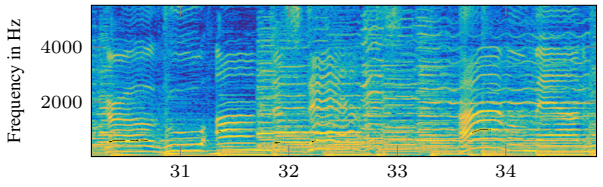
Although the focus in this study is on the application of upmixing, the position index $\hat{\Psi}$ can also be used for general spatial source separation applications. First tests gave appealing results which are comparable to [15]. In combination with the separation of direct and ambient signals further applications as center channel



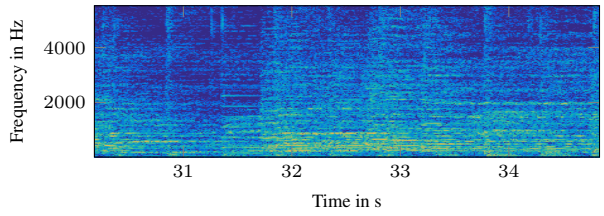

(a) *Left stereo channel* $X_L$




(b) *Right stereo channel* $X_R$




(c) *Extracted direct signal* $\hat{D}$




(d) *Extracted ambience signal* $\hat{N}$

Figure 6: *Spectrograms of the input signal a-b) and the extracted direct c) and ambient signals d).*

extraction [7], stereo speech enhancement [16,17] or the correction of panning errors with non-standard loudspeaker setups [18] could be possible.

## 5. DISCUSSION

The authors created a stereo to five channel upmixing VST plugin to test the practical suitability and sound quality of the described method. The target speaker layout and signal flow follows the diagram in Fig. 5. Rendered audio examples can be found on the website of the department[1].

The algorithm is quite efficient and only utilizes 3.0 % of a single CPU core on an Intel Core i5-2320 3 GHz processor at a sample rate of 44100 Hz. The size of the STFT blocks is set to 2048 samples with a hop size of 512 samples between two consecutive transforms. Profiling the code reveals that the main load is caused by the 2 FFT and 5 iFFT calculations which are required for a 2-to-5 frequency-domain upmix. The actual extraction of the source

[1] `http://www2.hsu-hh.de/ant/webbox/audio/kraft/DAFX15_upmix/`

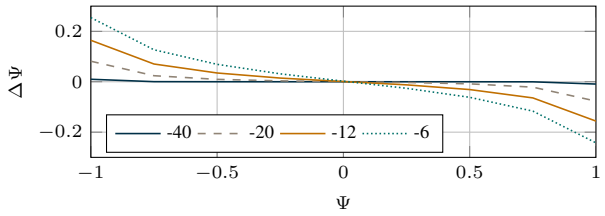



Figure 7: *Position index error* $\Delta\Psi$ *influenced by different ambience/direct power ratios in dB.*

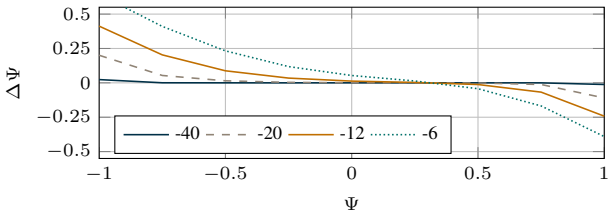



Figure 8: *Position index error* $\Delta\Psi$ *influenced by a second source with different power ratios (in dB) and fixed* $\Psi = 0.33$.

directions and the direct/ambience separation in the frequency domain with equations (14)-(15) and (19)-(20) only require a few instructions.

Spectrograms of the extracted signals for a sample song are shown in Fig. 6. It is apparent that the extracted direct signal in Fig. 6 c) includes most of the tonal energy from the input signal shown in Fig. 6 a-b). In contrast, the ambient signal Fig. 6 d) is diffuse and of lower energy. The tonal structure is only barely visible. The extracted ambient and direct signals sound realistic and although in particular the isolated ambient signal is not free of artefacts, they are not audible in the overall mix. This is caused by the fact that the upmixing process just redistributes signal components to more loudspeakers but nothing is removed or added compared to the original stereo signal. Although the positions of the sources are retained quite well when switching from stereo to surround, the perceived width of the spatial image tends to narrow a little bit in the upmix.

### 5.1. Estimated position error analysis

The generally proper reproduction of the source positions and pleasing ambient signals are a good indication that the simplifications and assumptions described in the above derivation do not cause any audible impairments and are valid for typical music material. However, it would be interesting to see how a violation of these assumptions will influence the accuracy of the estimated positions. For that purpose a direct signal consisting of white noise was panned to various positions $\Psi \in [-1, 1]$ and ambience was added using a *Large Hall* impulse response from a Bricasti M7 stereo reverb unit. The power ratio between the wet and dry signal was varied between $-40$ dB and $-6$ dB. The resulting error $\Delta\Psi = \hat{\Psi} - \Psi$ is plotted in Fig. 7 and it can be observed that the position error increases while the ambient signal power is increased. The consequence is that in particular extremely panned sources are estimated too close to the center. The same effect appears when a second source is added and the assumption of a single source per frequency band is violated. In Fig. 8 a white noise source signal

moving from left to right was overlaid with another fixed noise source at $\Psi = 0.33$ and different power levels. No ambient signal was added in this case. The error curve becomes asymmetric as the second source shifts the energy towards the right side but it has the same shape and behaviour as in the previous example. Still the error increases with a higher power of the disturbing source and with increased difference between both source positions. Although this is no detailed error analysis and only synthetic signals were used, the results confirm and visualise what was already perceived in the upmix application described before.

### 5.2. STFT resolution

The trade-off between time and frequency resolution is an important parameter for the algorithm as it is only capable to deal with a single source in a specific time-frequency point $(b, k)$. Different block lengths for the STFT in a range from 256 to 8192 samples were investigated at a sample rate of 44100 Hz. The best sounding results are achieved with block lengths of 2048 samples, whereas the difference to 1024 or 4096 samples is only barely audible. First attempts were made to summarise frequency bins in the STFT into perceptual bands as done by Faller [5]. No obvious artefacts appeared even if the 2048 spectral bins were summarised in only 24 Bark bands. This opens a wide field of different methods to reduce and interpolate the spectral resolution and it might be in particular interesting to see if this enables us to increase the precision of the position estimation by using redundant information.

## 6. CONCLUSION

In this paper a new algorithm to estimate the perceived azimuth directions in a stereo signal was derived from a typical signal model. With the estimated directions it is possible to separate the direct and ambient signal components and to remix the stereo signal. Both, the signal separation and estimation of directions show similarity to a classical mid-side decomposition of stereo signals. However, in the presented form it is applied in sub-bands with the help of a short-time fourier transform and generalised to non-center panned signals. This allows to separate multiple sources with the constraint that only one dominant source is active at a specific time instant and frequency band. An implementation as a stereo to five channel upmixing VST plugin demonstrates the applicability and high sound quality of the proposed method at a very low computational cost.

For future enhancements it would be interesting to further investigate the influence of the STFT resolution and in particular the usage of perceptually motivated non-linear resolutions on the quality of the separated signals. Another aspect becoming more important with an increasing amount of loudspeakers is a proper decorrelation of the ambient signals to achieve a smooth and diffuse ambient sound field. Several approaches are known in the literature but their suitability in the context of the presented upmix algorithm have not been examined by the authors, yet.

## 7. REFERENCES


[1] Michael A. Gerzon, "Optimal Reproduction Matrices for Multispeaker Stereo," in *Proc. of the 91st AES Convention*, 1991.

[2] Roger Dressler, "Dolby Surround Pro Logic II decoder principles of operation," *Dolby White paper*, 2000.

[3] Carlos Avendano and Jean-Marc Jot, "A frequency-domain approach to multichannel upmix," *Journal of the Audio Engineering Society*, vol. 52, no. 7, pp. 740–749, 2004.

[4] Carlos Avendano and Jean-Marc Jot, "Ambience extraction and synthesis from stereo signals for multi-channel audio up-mix," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2002.

[5] Christof Faller, "Multiple-loudspeaker playback of stereo signals," *Journal of the Audio Engineering Society*, vol. 54, no. 11, pp. 1051–1064, 2006.

[6] Michael M. Goodwin and Jean-Marc Jot, "Primary-ambient signal decomposition and vector-based localization for spatial audio coding and enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007.

[7] Earl Vickers, "Frequency-Domain Two- to Three-Channel Upmix for Center Channel Derivation and Speech Enhancement," in *Proc. of the 127th AES Convention*, 2009.

[8] John Usher and Jacob Benesty, "Enhancement of Spatial Sound Quality: A New Reverberation-Extraction Audio Upmixer," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2141–2150, Sept. 2007.

[9] Alexander Jourjine, Scott Rickard, and Özgür Yilmaz, "Blind separation of disjoint orthogonal signals: demixing N sources from 2 mixtures," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2000.

[10] Roger Dressler, "Dolby Surround Pro Logic Decoder Principles Of Operation," *Dolby White paper*, 1982.

[11] Benjamin B. Bauer, "Phasor analysis of some stereophonic phenomena," *IRE Transactions on Audio*, vol. 10, no. 1, pp. 143–146, 1962.

[12] Ville Pulkki, "Virtual sound source positioning using vector base amplitude panning," *Journal of the Audio Engineering Society*, vol. 45, no. 6, pp. 456–466, 1997.

[13] Gary S. Kendall, "The Decorrelation of Audio Signals and Its Impact on Spatial Imagery," *Computer Music Journal*, vol. 19, no. 4, pp. 71–87, 1995.

[14] Marco Fink, Sebastian Kraft, and Udo Zölzer, "Downmix-compatibe conversion from mono to stereo in time- and frequency-domain," in *Proc. of the 18th Int. Conference on Digital Audio Effects*, 2015.

[15] Dan Barry, Bob Lawlor, and Eugene Coyle, "Sound source separation: Azimuth discriminiation and resynthesis," in *Proc. of the 7th Int. Conference on Digital Audio Effects*, 2004.

[16] Alexandra Cracuin, Christian Uhle, and Tom Bäckström, "An evaluation of stereo speech enhancement methods for different audio-visual scenarios," in *Proc. of the 23rd European Signal Processing Conference (EUSIPCO)*, 2015.

[17] Jürgen T. Geiger, Peter Grosche, and Yesenia Lacouture Parodi, "Dialog Enhancement of Stereo Sound," in *Proc. of the 23rd European Signal Processing Conference (EUSIPCO)*, 2015.

[18] Alexander Adami, Michael Schoeffler, and Jürgen Herre, "Re-Panning of Directional Signals and its Impact on Localization," in *Proc. of the 23rd European Signal Processing Conference (EUSIPCO)*, 2015.