TIME-DOMAIN IMPLEMENTATION OF A STEREO TO SURROUND SOUND UPMIX ALGORITHM

Sebastian Kraft, Udo Zölzer

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

ABSTRACT

This paper describes a time-domain algorithm to upmix stereo recordings for an enhanced playback on a surround sound loudspeaker setup. It is mainly the simplified version of a previously published frequency-domain algorithm where the standard short-time Fourier transform is now replaced by an IIR filter bank. The design of complementary filter blocks and their arrangement in a tree structure to form a filter bank are derived. The arithmetic complexity of the filter bank itself and of the complete upmix algorithm is analysed and compared to the frequency-domain approach. The time-domain upmix is less flexible in its configuration but achieves an audio quality comparable to the frequency-domain implementation at a fraction of its computational cost.

1. INTRODUCTION

Upmixing is a process to generate additional channels when an audio source is intended to be played back over a setup with more loudspeakers than source channels. Ideally, a good upmix should redistribute the input signal to all available loudspeakers providing an immersive listening experience without compromising the original character of the stereo track. For example, the azimuth position of sources in a stereo mix as well as the overall timbre and spatial character should be preserved.

Most algorithms in this area are based on the same processing principles applied to a time-frequency-domain representation of the input signal. The channels of a stereo or multi-channel recording are described as a weighted sum of direct signal sources overlaid by an uncorrelated ambient signal [1, 2]. First, the azimuth positions or panning coefficients of the sources are estimated under the assumption that only one dominant source is active at a single time-frequency instant. Next, the direct and ambient components are separated with the knowledge of the panning coefficients. Having the separated signals with their related source positions it is possible to remix the original content considering any different target loudspeaker configuration.

Other algorithms mainly focus on a direct and ambient signal separation [3, 4], where [3] is one of the few examples for a time-domain approach. However, as a normalised least mean squares (NLMS) method is used to adapt an FIR extraction filter with several hundreds of coefficients it is computationally quite demanding. Further examples for time-domain approaches are Dolby Pro Logic I and II [5] which were widely spread in the consumer area a decade ago and were optimised for a cost effective implementation using simple time-domain operations. Basically, only a few subtractions and additions of the left and right channels with additional phase shifts and VCAs (voltage controlled amplifiers) for simple directional steering are required. But due to full-band processing of the input signal the capability to separate multiple concurrent sources is quite limited.

The idea behind the approach presented in this paper and its derivation is similar to the previous work in [6, 7] but all derivations are rewritten to yield an equivalent time-domain formulation. This allows to replace the short-time Fourier transform (STFT), previously used to create a time-frequency representation of the input signal, by a non-subsampling filter bank as in [8]. It is build of complementary IIR filters arranged in a tree structure and allows a perfect allpass reconstruction. In [2, 7] it was already pointed out that the STFT spectra resolution could be reduced to Bark bands without impairments. Hence, a low-resolution filter bank would be well suited to search for an optimal trade-off between complexity and frequency resolution and to analyse the influence of the time-frequency resolution on the quality of the resulting upmix.

The underlying stereo signal model and estimation of source positions as well as the direct and ambient component separation will be introduced in Sec. 2. Afterwards, the design of the filter bank is described in Sec. 3 together with an analysis of its arithmetic complexity. Section 4 shows the application of the proposed separation in the context of a stereo to multi-channel surround sound upmix and compares the results with its frequency-domain counterpart.

2. SIGNAL SEPARATION

2.1. Stereo signal model

The left and right channels of a stereo signal

$$x_L(n) = \left[\sum_{i=1}^{I} g_{L_i} \cdot s_i(n)\right] + a_L(n) \tag{1}$$

$$x_R(n) = \left[\sum_{i=1}^{I} g_{R_i} \cdot s_i(n)\right] + a_R(n) \tag{2}$$

can be described as a weighted sum of I source signals $s_i(n)$ and additive uncorrelated ambient signals $a_L(n)$ and $a_R(n)$ in the left and right channel, respectively. The weightings g_{L_i} and g_{R_i} of the individual sources are called panning coefficients and are bound between zero and one. Their squared sum should be equal to one ($g_{L_i}^2 + g_{R_i}^2 = 1$) to achieve a constant power and loudness of a panned source independent of its current position. By applying a

filter bank, the signal model

$$x_L(b,k) = \left[\sum_{i=1}^{I} g_{L_i} \cdot s_i(b,k)\right] + a_L(b,k) \tag{3}$$

$$x_R(b,k) = \left[\sum_{i=1}^{I} g_{R_i} \cdot s_i(b,k)\right] + a_R(b,k) \tag{4}$$

is transformed into a time-frequency representation. The variables b and k denote the time and band index, whereas b is equal to n for non sub-sampling filter banks and will be used in the following.

Two simplifications are required to invert the signal model and to recover sufficient approximations of the source signals and their panning parameters. First, it is a typical assumption that at a certain time instant n and in a frequency band k only a single dominant source s_u is active and the contribution of other sources is close to zero ($\sum_{\forall i \neq u} |s_i(n,k)| \approx 0$) if the time and frequency resolution is not too small [9]. This allows to summarise the time-frequency representations of the individual sources

$$g_L(n,k) \cdot s(n,k) \approx \sum_{i=1}^{I} g_{L_i} \cdot s_i(n,k) \tag{5}$$

into a single source $s(n,k)$ and panning coefficients $g_{L/R}(n,k)$.

Second, the left and right ambient signal can be expected to sound similar but due to different paths and reflections in the room, they are decorrelated. Hence, the left and right ambient signals

$$a_L(n) = \mathcal{H}_{A_L}\{a(n)\}, \qquad a_R(n) = \mathcal{H}_{A_R}\{a(n)\}$$

originate from a single ambient signal $a(n)$ which has been modified by an abstract filter operation $\mathcal{H}(\cdot)$. Combining both assumptions, a simplified time-frequency signal model

$$x_L(n,k) = g_L(n,k) \cdot s(n,k) + \mathcal{H}_{A_L}\{a(n,k)\}, \tag{6}$$

$$x_R(n,k) = g_R(n,k) \cdot s(n,k) + \mathcal{H}_{A_R}\{a(n,k)\} \tag{7}$$

with a reduced number of unknowns can be obtained.

If a STFT would be used for the time-frequency transform, the abstract filter operation in the signal model could be implemented as a multiplication of $A(n,k)$ with a frequency response

$$H_A(k) = \gamma(k) \cdot e^{j\phi(k)}, \qquad \begin{aligned} 0 < \gamma(k) &< 1 \\ 0 < \phi(k) &\leq \pi \end{aligned} \tag{8}$$

consisting of a band-wise magnitude $\gamma(k)$ and phase $\phi(k)$ coefficient. This leads to a signal model

$$X_L(b,k) = g_L(b,k) \cdot S(b,k) + H_{A_L}(k) \cdot A(b,k) \tag{9}$$

$$X_R(b,k) = g_R(b,k) \cdot S(b,k) + H_{A_R}(k) \cdot A(b,k) \tag{10}$$

in the time-frequency domain. The actual decorrelation filter parameters are usually not known for general music material and are difficult to estimate. However, a coarse approximation of a decorrelation filter response by a random distribution can lead to realistically sounding decorrelated signals [10, 11, 12] and was also successfully used for ambience extraction in [7]. Furthermore, the desired correlation of the left and right ambient signal can be nicely adjusted by the phase angle ϕ where $\phi = \pi/2$ would yield a broadband correlation of 0 and with $\phi = \pi$ the resulting ambient signals are out of phase (correlation is -1).

With a time-domain filter bank, the application of the ambience decorrelation filters $\mathcal{H}_{A_L}$ and $\mathcal{H}_{A_R}$ is not as trivial as in the frequency-domain. An equivalent time-domain formulation of the filter in (8) with a corresponding signal model could look like

$$h_{a_{L/R}}(k) = \gamma(k) \cdot \pm 1, \qquad 0 < \gamma(k) < 1 \tag{11}$$

$$x_L(n,k) = g_L(n,k) \cdot s(n,k) + h_{a_L}(k) \cdot a(n,k) \tag{12}$$

$$x_R(n,k) = g_R(n,k) \cdot s(n,k) + h_{a_R}(k) \cdot a(n,k) \tag{13}$$

where the filter is modelled with a band-wise gain $\gamma(k)$ and a phase shift of π can be achieved by choosing opposite sign coefficients in each channel. Other phase shifts would require a time-domain filtering of each sub-band and would impose the problem of inverse filtering when extracting the direct and ambient signal in Sec. 2.3.

2.2. Estimation of source directions

For typical music mixes the amplitude of the ambient signal $a(n,k)$ can be assumed to be far less than the amplitude of the direct signal $s(n,k)$. This also means that the power of the left and right channels

$$P_{x_L}(n,k) \approx g_L^2(n,k) \cdot P_s(n,k) \tag{14}$$

$$P_{x_R}(n,k) \approx g_R^2(n,k) \cdot P_s(n,k). \tag{15}$$

is mainly depending on the weighted direct signal power. Rearranging and solving equations (14)-(15) with the constraint $g_L^2 + g_R^2 = 1$, the panning coefficients

$$\hat{g}_L(n,k) = \sqrt{\frac{P_{x_L}(n,k)}{P_{x_L}(n,k) + P_{x_R}(n,k)}} \tag{16}$$

$$\hat{g}_R(n,k) = \sqrt{\frac{P_{x_R}(n,k)}{P_{x_L}(n,k) + P_{x_R}(n,k)}}. \tag{17}$$

can be estimated from the power of the left and right stereo channels. A simple estimate of the power of a sub-band signal $x(n,k)$

$$P_x(n,k) = \alpha \cdot P_x(n-1,k) + (1-\alpha) \cdot x(n,k)^2 \tag{18}$$

can be determined by recursive averaging with a coefficient $0 < \alpha < 1$. Other power estimates, in particular with signal adaptive coefficients, may be investigated in the future and could for example improve processing of transients.

The "stereophonic law of sines" [13]

$$\frac{g_L - g_R}{g_L + g_R} = \frac{\sin(\theta)}{\sin(\theta_0/2)} = -\Psi \tag{19}$$

describes the perceived angle θ of a source if its amplitude is weighted by $g_{L/R}$ for playback on a left and right loudspeaker with an angle θ_0 between both. The normalised position index Ψ, ranging from -1 for left and $+1$ for right positions, combines the coefficients $g_{L/R}$ in a single value. From (14)-(15) and (19) one can derive estimates for the position index and angle

$$\hat{\Psi}(n,k) = \frac{\sqrt{P_{x_R}(n,k)} - \sqrt{P_{x_L}(n,k)}}{\sqrt{P_{x_R}(n,k)} + \sqrt{P_{x_L}(n,k)}} \tag{20}$$

$$\hat{\theta} = \arcsin\left(\sin(\theta_0/2) \cdot \hat{\Psi}(n,k)\right) \tag{21}$$

based on the power of the left and right stereo channel.

2.3. Direct and ambient signal separation

When the panning coefficients or their estimates from the previous section are known, the signal model (12)-(13) can be transformed mathematically to get the direct and ambient signal components

$$\hat{s}(n,k) = \frac{h_{a_R}(k)\cdot x_L(n,k) - h_{a_L}(k)\cdot x_R(n,k)}{h_{a_R}(k)\cdot \hat{g}_L(n,k) - h_{a_L}(k)\cdot \hat{g}_R(n,k)} \tag{22}$$

$$\hat{a}(n,k) = \frac{\hat{g}_L(n,k)\cdot x_R(n,k) - \hat{g}_R(n,k)\cdot x_L(n,k)}{\hat{g}_L(n,k)\cdot h_{a_R}(k) - \hat{g}_R(n,k)\cdot h_{a_L}(k)}\,. \tag{23}$$

For low-resolution filter banks as used in the upmix application the random decorrelation filter gains $\gamma(k)$ from (11) would cause an audible band-wise panning instead of the desired diffuse decorrelation. Hence, the decorrelation filters are set to

$$h_{a_L}(k) = 1, \qquad h_{a_R}(k) = -1 \tag{24}$$

and in this case the extraction formula can be further simplified to

$$\hat{s}(n,k) = \frac{x_L(n,k) + x_R(n,k)}{\hat{g}_L(n,k) + \hat{g}_R(n,k)} \tag{25}$$

$$\hat{a}(n,k) = \frac{\hat{g}_L(n,k)\cdot x_R(n,k) - \hat{g}_R(n,k)\cdot x_L(n,k)}{\hat{g}_L(n,k) + \hat{g}_R(n,k)}\,. \tag{26}$$

As already pointed out in [6], the above equations are very similar to a classical mid-side decomposition performed in sub-bands. The main difference is the weighting with the estimated panning coefficients to allow proper separation of ambient and direct components in case the direct signal is not panned to the center.

3. COMPLEMENTARY FILTERBANK

An IIR filter bank as in [8, 14] is used to create a time-frequency representation of the input signal. It consists of complementary filter blocks arranged in a tree structure and additional allpass sections are inserted to achieve an overall allpass reconstruction property. The individual bands will not be downsampled, hence the reconstruction can be a simple summation of the sub-bands.

3.1. Complementary allpass filter structure

The basic building block of the filter bank is a filter pair F and $\tilde{F}$ which split an input signal into a lower and higher complementary band. The two filters are power complementary if their transfer functions satisfy

$$\left|F(e^{j\omega})\right|^2 + \left|\tilde{F}(e^{j\omega})\right|^2 = 1 \tag{27}$$

and if the sum of the transfer functions additionally yields an allpass magnitude response

$$\left|F(e^{j\omega}) + \tilde{F}(e^{j\omega})\right| = 1 \tag{28}$$

they are *doubly complementary* [15]. Therefore, by summation of both filter outputs it is possible to recover the input signal except for a certain phase shift.

The pair of filters could be directly designed with standard methods allowing for above constraints. However, by decomposing a single filter prototype F in two parallel allpass sections it is possible to obtain a second complementary output with just one further subtraction. The detailed derivation can be found in [16] and will be shortly summarised in the following sections.

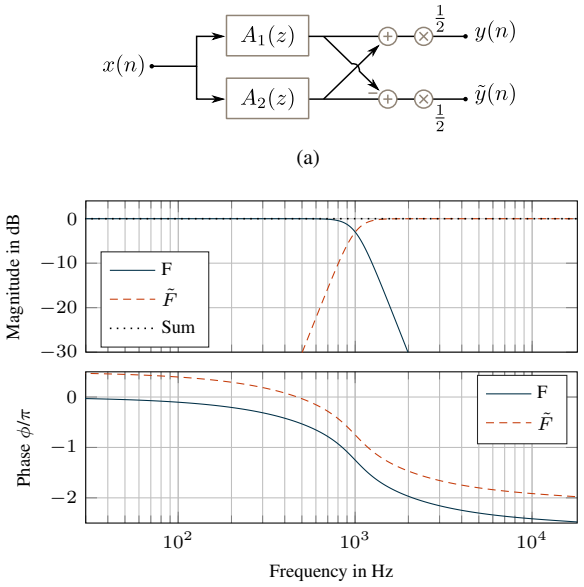


Figure 1: *Doubly complementary allpass filter structure (a) with exemplary magnitude and phase response for a 5th order butterworth lowpass (b).*

3.1.1. Allpass Decomposition

An IIR filter $F(z)$ can be split into a parallel sum

$$F(z) = \frac{P(z)}{D(z)} = \frac{1}{2}\Big(A_1(z) + A_2(z)\Big) \tag{29}$$

of two allpass filters $A_1(z)$ and $A_2(z)$ in case the following requirements are met:

1. the order N of $F(z)$ is odd
2. $P(z)$ is a mirror symmetric polynomial, $P(z^{-1}) = z^N\,P(z)$
3. $F(z)$ has real coefficients and $|F(e^{j\omega})| \leq \infty$ (*bounded real* transfer function)

This holds true for typical IIR filter designs like Butterworth, Chebyshev and elliptic filters. Furthermore, a complementary filter

$$\tilde{F}(z) = \frac{Q(z)}{D(z)} = \frac{1}{2}\Big(A_1(z) - A_2(z)\Big) \tag{30}$$

can be easily obtained by the difference of the allpass filters as depicted in Fig. 1. The summed transfer function of a complementary filter stage formed by (29) and (30)

$$H(z) = F(z) + \tilde{F}(z) = A_1(z) \tag{31}$$

has the required allpass characteristic. This means that perfect magnitude reconstruction is possible but the phase shift and group delay of $A_1(z)$ will remain in the reconstructed signal.

The respective order of the allpass sections is $N_1 = (N-1)/2$ and $N_2 = (N+1)/2$ where $N_1 + N_2 = N$. Table 1 gives an overview of the required instructions per sample for a complementary allpass filter structure compared to a direct form implementation with two separate filters. It can be seen that the complementary filter created with the allpass decomposition requires only little more than half of the instructions as a direct-form implementation with two filters.

	Mult.	Add.	Overall
Compl. Allpass struct.	$N+2$	$2N+2$	$3N+4$
2x Direct-Form filters	$3N+1$	$4N$	$7N+1$

Table 1: *Number of multiplications, additions and overall operations for a single complementary filter stage per sample.*

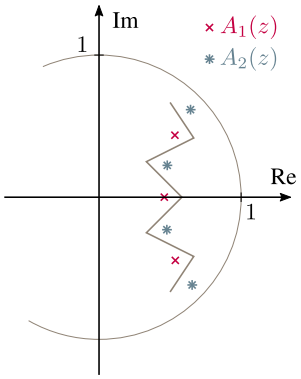


Figure 2: *Alternating selection of poles in the z-plane to create an allpass decomposition.*

3.1.2. Filter design

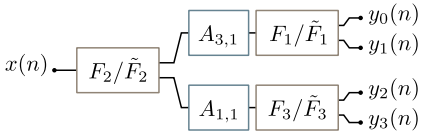
The general form of the two allpass filters

$$A_1(z) = \prod_{k=1}^{N_1} \frac{z^{-1}-p_k}{1-z^{-1}\,p_k}$$
$$A_2(z) = \prod_{k=N_1+1}^{N} \frac{z^{-1}-p_k}{1-z^{-1}\,p_k} \qquad (32)$$

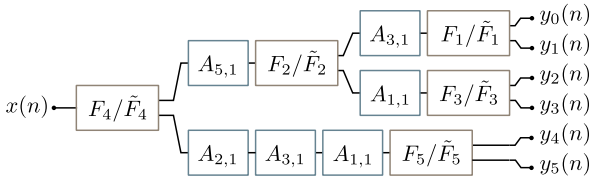
is fully defined by the knowledge of the poles p_k, as the zeros $z_k = 1/p_k$ are the inverse of the poles. From (29) and (30) it is apparent that $F(z)$ or $\tilde{F}(z)$ feature the same poles as the sum or difference of $A_1(z)$ and $A_2(z)$. Hence, the poles of the allpass filters are just subgroups of the poles already contained in $F(z)$.

An algorithm to derive a suitable grouping of the poles from a filter transfer function $F(z)$ is described in [16] where first the polynomial $Q(z)$ is determined and then the zeros of $P(z)+Q(z)$ are calculated. The zeros outside the unit circle will form the first allpass $A_1(z)$ and the zeros inside the unit circle will form the second allpass $A_2(z)$. The algorithm to calculate $Q(z)$ is recursive and requires several polynomial multiplications which may become numerically unstable for high order filters. In particular Butterworth designs lead to very small valued coefficient sets and it may become difficult to find a stable decomposition for filter orders above 5.

A more simple graphical allpass decomposition is given in [17] which directly makes use of the poles typically derived in the IIR design procedure and hence is less prone to numerical errors. The poles of Butterworth, Chebyshev and elliptic lowpass filters are placed on an ellipsoid curve inside the unit circle. An alternating separation as depicted in Fig. 2 yields two groups of poles, whereas the smaller group with $(N-1)/2$ poles is assigned to $A_1(z)$ and the remaining $(N+1)/2$ poles are assigned to $A_2(z)$.



(a) $M = 3$ filters, 4 output bands



(b) $M = 5$ filters, 6 output bands

Figure 3: *Two exemplary filter bank structures consisting of 3 and 5 complementary filters.*

3.2. Filter bank tree structure

In the following, $H_k(z) = Y_k(z)/X(z)$ will denote the transfer function of a filter bank channel k. Overall M complementary filters $F_m(z)/\tilde{F}_m(z)$ are used and $A_{m,1}(z)$ and $A_{m,2}(z)$ are the corresponding allpass decompositions.

The complementary filter stages are cascaded in a tree structure to successively divide the bands and to create $M+1$ outputs. An exemplary filter bank with 4 outputs and 3 complementary filters is shown in Fig. 3 a). The additional allpass sections $A_{3,1}(z)$ and $A_{1,1}(z)$ are required to guarantee an overall allpass transfer function after summing the sub-band outputs for reconstruction. For example, when summing the individual output transfer functions without the allpass sections it appears that

$$\begin{aligned} H(z) &= H_0(z) + H_1(z) + H_2(z) + H_3(z) \\ &= F_2(z)\cdot\Big(F_1(z)+\tilde{F}_1(z)\Big) + \tilde{F}_2(z)\cdot\Big(F_3(z)+\tilde{F}_3(z)\Big) \\ &= F_2(z)\cdot A_{1,1}(z) + \tilde{F}_2(z)\cdot A_{3,1}(z), \end{aligned}$$

is not an allpass. However, by adding additional allpass sections after $F_2/\tilde{F}_2$ the overall transfer function

$$\begin{aligned} H(z) &= F_2(z)\cdot A_{3,1}(z)\cdot A_{1,1}(z) + \tilde{F}_2(z)\cdot A_{1,1}(z)\cdot A_{3,1}(z) \\ &= A_{2,1}(z)\cdot A_{1,1}(z)\cdot A_{3,1}(z) \end{aligned}$$

becomes allpass. Another exemplary filter bank structure to yield six doubly complementary bands is given in Fig. 3 b). If a reconstruction is not required in the desired application and the filter bank is only used for signal analysis, the additional allpass sections can be omitted without altering the power of the sub-band signals.

More details about the general setup of tree-structured recursive filter banks and in particular about the placement of the additional allpass sections to yield allpass reconstruction properties can be found in [14]. In a summary, the basic rules are:

- The overall transfer function for a bank of M filters is

$$H(z) = \sum_{k=0}^{M} H_k(z) = \prod_{m=1}^{M} A_{m,1}(z). \qquad (33)$$

	Mult.	Add.	Overall
Compl. filters	$N+2$	$2\,N+2$	$M\cdot(3\,N+4)$
Add. allpass	N_1	$2\,N_1$	$M_A\cdot 3\,N_1$

Table 2: *Number of multiplications, additions and overall operations per sample for a filter bank with M filters of order N. M_A denotes the required number of allpass sections for perfect magnitude reconstruction.*

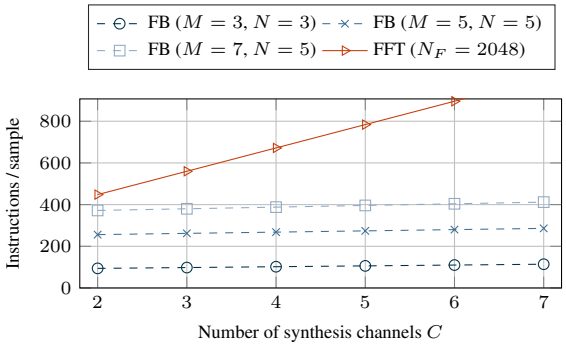


Figure 4: *Arithmetic complexity for filter banks with 2 analysis and C synthesis channels compared to an STFT analysis/synthesis.*

Hence, the overall group delay only depends on the $A_{m,1}(z)$ sections and can be minimised by always applying the lower number of poles to $A_{m,1}(z)$ during the allpass decomposition.

- It has to be assured that every sub-band signal passes all possible allpass sections ($A_{m,1}(z)$, $m=[1\ldots M]$) on its way through the filter bank. Missing allpass sections have to be inserted into the signal path.
- The required number of additional allpass sections can be minimised with a simple rule: At every branch we have to add the corresponding allpass sections of all filters in the opposite branch.

3.3. Arithmetic complexity

The arithmetic complexity in terms of required multiplications and additions for a filter bank with M filters of order N is given in Table 2. It can be seen that the complexity linearly increases with the number of bands and filter order.

The STFT is the standard transform for analysis/synthesis systems and the question is how it would compare to a filter bank based approach in terms of arithmetic complexity. The number of operations to transform a block of length N_F with a standard Cooley-Tukey FFT (radix-2) can be estimated to be in the range of $\sim\ 5\log(N_F)\,N_F$. With a typical overlap of 75 % and real valued time-domain signals this yields $\sim 10\log(N_F)$ instructions per sample and transform.

Figure 4 compares several recursive filter banks and a STFT variant with a block size of $N_F=2048$ samples. The complexity of the filter bank analysis/synthesis system is mainly independent of the number of output channels as the synthesis is a simple summation of all sub-bands. In contrast for the STFT, as well as sub-sampled filter banks, an equal complexity synthesis step is required and the complexity increases linearly with the number of synthesis channels. It can be seen that a filter bank may be in particular advantageous if the application only requires a few bands with low filter orders but arbitrary frequency resolution and if more output than input channels are to be generated.

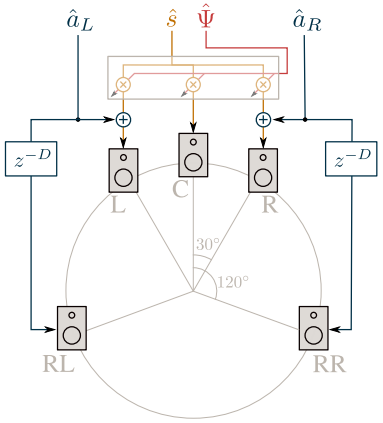


Figure 5: *Stereo to 5 channel upmix.*

	M	M_A	N	Frequencies [Hz]
FB I	3	2	3	220, 1000, 4000
FB II	5	6	3	220, 500, 1000, 2000, 5000
FB III	5	6	5	220, 500, 1000, 2000, 5000
STFT	$N_F=2048$, $N_H=512$			

Table 3: *Chosen filter bank parameters.*

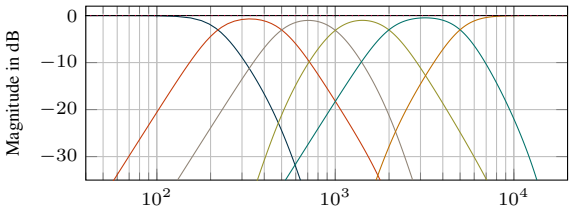
4. UPMIX APPLICATION

The estimated direct signal source positions and the separated direct and ambient signals can be used to create a stereo to surround upmix following a signal flow as depicted in Fig. 5. The direct signal is repanned on the front loudspeakers, for example by using *Vector Base Amplitude Panning* (VBAP) [18], while the ambient signals are added to the corner loudspeakers. To decorrelate the front from the rear ambient signals, a short delay is included but it would also be possible to apply more advanced time-domain decorrelators as described in [19].

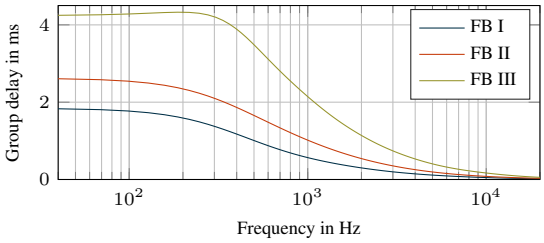
4.1. Filter bank configuration

Several combinations of filter order as well as corner frequencies and number of bands were tested and the corresponding parameters are given in Table 3. The magnitude response for configuration FB II is depicted in Fig. 6 a) and the group delay for all given configurations is plotted in Fig. 6 b). The coefficient α for the recursive power estimation per sub-band was set to $\alpha=5\cdot 10^{-4}$ in the following experiments.

It turned out, that the positions estimated from the lowest and highest band of the filter bank are not reliable as there is too much overlap of individual sources in these frequency regions. Therefore, in the following only filter bank outputs 1 up to $M-1$ will

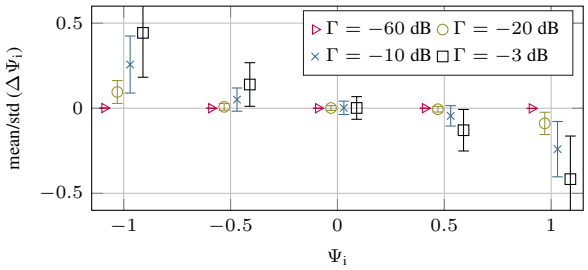


(a) Magnitude response of FB II

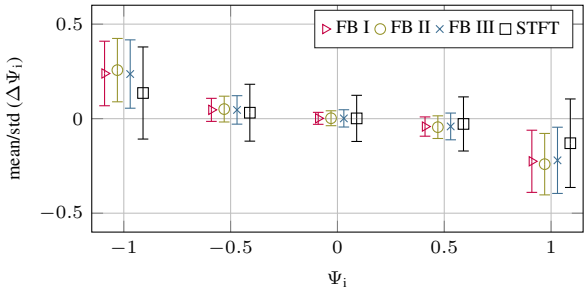


(b) Group delay

Figure 6: *Magnitude response and group delay for several filter bank configurations.*



(a) Filter bank FB II



(b) Comparison of filter banks and STFT at $\Gamma = -10$ dB

Figure 7: *Mean and standard deviation of $\Delta\Psi_i$ for various configurations.*

be processed by the upmix and channel 0 and M will be directly fed to the front left and right loudspeakers. In the end, this corresponds to a band-limiting of the extracted ambient and center channels.

4.2. Evaluation of estimated positions

It was assumed in Sec. 2.2 that a low power ambient signal will not influence the estimation of the panning coefficients. However, the question is how the accuracy of the panning estimation is impaired if the ambient signal power is increased. To further investigate this, a single direct signal has been panned to various positions Ψ_i and ambience was added with a *Large Hall* impulse response from a *Bricasti M7* stereo reverb unit. The ambient to direct power ratio (ADR)

$$\Gamma = 10\ \log_{10}\left(\frac{P_{A_L} + P_{A_R}}{P_S}\right) \tag{34}$$

describes the ratio between the overall ambient and direct signal power where

$$P_x = \sum_n x(n)^2$$

denotes power of a signal $x(n)$. The resulting mean and standard deviation of the position error

$$\Delta\Psi_i(n,k) = \hat{\Psi}_i(n,k) - \Psi_i \tag{35}$$

is plotted for several ADR values in Fig. 7 a) and it can be seen that stronger ambient components directly lead to a higher error. As the ambient signal has near equal power in the left and right channel the estimated positions will be biased towards the center and the error further increases for strongly panned sources. In Fig. 7 b) the error for $\Gamma = -10$ dB is compared between the different filter bank configurations from Table 3 and also to the STFT based method from [6]. It is apparent that the error is relatively independent of the filter bank configuration. Compared to the STFT based position estimation the mean error for strongly panned sources is higher, however, the standard deviation is considerably lower for all source positions.

4.3. Ambient signal quality

Although the extracted ambient signals sound realistic they are too different from the real signals which were added in the mixing process and a direct comparison, e.g. by error energy, is usually not significant. Therefore, only signal characteristics can be compared and it is well known that ambient signals should be diffuse and decorrelated to create an immersive sound field. The diffuseness and uniformity of an ambient signal can be measured by inter-channel cross-correlation (ICC) and inter-channel level differences (ICLD), whereas for real ambient signals both are observed to be close to zero. With a high-resolution frequency-domain implementation it is possible to apply versatile decorrelation filters with little effort as described in detail in [7]. With a low-resolution filter bank this is not feasible and with the simple filters as in (24) the left and right ambient signals are just phase-inverse copies of each other. Hence, the ICC is -1 and the ICLD is 0. Listening to the isolated ambient sound field, the out of phase character creates an impression of width but it can also evoke unpleasant cancellation artefacts in particular during head movements. This is the main difference to the frequency-domain implementation [7] where the resulting ICC was freely adjustable.

4.4. Arithmetic complexity

A detailed analysis of the arithmetic complexity of the frequency-domain upmix algorithm was done in [7] and the numbers in Table 4 a) are based on these findings. However, for a fair comparison with the simplified time-domain approach, the ambient decor-

relation filters in the frequency-domain were set to $H_{A_L}(k) = 1$ and $H_{A_R}(k) = -1$ according to (24) and no further decorrelation between front and rear channels is applied. This partly reduces the arithmetic complexity of the direct and ambient decomposition.

A corresponding analysis of the arithmetic complexity of the filter bank based upmix algorithm is given in Table 4 b). As expected, the base cost for analysis and synthesis is drastically reduced for typical filter bank configurations. In contrast, the number of operations required for the upmix processing has increased as all sub-bands are processed at full rate. Processing an upmix from stereo to 5-channel surround in the frequency-domain requires about 38 MFlops, whereas the corresponding time-domain variant with a low-resolution filter bank is in a range between 8 and 20 MFlops. This could be further reduced by a lowered update rate of the estimated positions and repanning coefficients but was out of the scope of this study due to the large amount of possible solutions and parameters.

It has to be noted that highly optimised FFT implementations (e.g. the FFTW library[1]) are widely available and can easily reduce the processing time for a transform by a factor of three up to five. Of course, similar optimization strategies like code vectorisation could be applied to the filter bank. But as no ready to run solutions are available this would require a comparably high effort and programming expertise.

4.5. Discussion

Informal listening tests confirm that the generated 5-channel surround upmix is convincing and works well with typical commercial studio music recordings. Compared to the stereo input, the source positions are well retained and no timbral coloration or other artefacts are audible. The center loudspeaker successfully stabilises the front image, in particular for listeners outside of the sweet spot and the out of phase artefacts observed with the isolated ambient signal are masked by the usually quite strong direct signal and are not annoying. However, in a direct A/B comparison with the STFT based frequency-domain upmix it is possible to spot subtle differences. The ambient signal is weaker and does not create a sound field as diffuse as experienced from the STFT approach.

First tests with discrete stereo microphone recordings showed good results for coincident microphone arrangements. In contrast, with non-conincident microphone setups a proper estimation of directions and a separation of direct and ambient components is not possible at the same quality. The reason is that phase shifts are introduced in the direct signals between both channels violating the basic signal model assumptions in Sec. 2.1.

Overall, the time-domain implementation upmix offers obvious benefits compared to a simple stereo playback even with the lowest resolution filter bank FB I. The actual configuration of the filter bank does not seem to have a strong influence on the results. More important seems the fact that low and high frequency content is separated by the filter bank and will not interfere with the estimation in the important mid frequency regions.

Another interesting aspect is the filter bank group delay of less than 5 ms which is quite low compared to the blocking delay of a STFT based analysis and synthesis. A STFT configuration with a block size $N_F = 2048$ at 44.1 kHz sample rate would yield about 46 ms delay. Therefore, the time-domain upmix is well suited for real-time applications and implementations on low-cost stream-based DSPs where no block-based processing is possible or FFT implementations are not available.

5. CONCLUSION

The goal of this study was to develop a low-complexity time-domain upmix algorithm. First, an equivalent time-domain formulation to a previously described frequency-domain method for estimation of source positions and separation of direct and ambient signal components has been derived. A filter bank is then used to create a time-frequency representation of the input signal and its design based on complementary IIR filters is described.

The arithmetic complexity of the filter bank and the filter bank-based upmix is compared to a STFT based approach. The time-domain variant is less flexible in its possible configurations but achieves an audio quality comparable to the frequency-domain approach at a fraction of computational cost. This makes it in particular well suited for low-cost and sample-by-sample DSP implementations where no highly optimised FFT implementation is available or for low-delay applications.

Sound examples of both the frequency-domain and time-domain approach can be found on the website of the department[2].

6. REFERENCES

[1] 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.

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

[3] 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, sep 2007.

[4] Jianjun He, Ee-Leng Tan, and Woon-Seng Gan, "Linear Estimation Based Primary-Ambient Extraction for Stereo Audio Signals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 505–517, feb 2014.

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

[6] Sebastian Kraft and Udo Zölzer, "Stereo signal separation and upmixing by mid-side decomposition in the frequency-domain," in *Proc. of the 18th Int. Conference on Digital Audio Effects*, 2015.

[7] Sebastian Kraft and Udo Zölzer, "Low-complexity stereo signal decomposition and source separation for application in stereo to 3D upmixing," in *Proc. of the 140th AES Convention*, 2016.

[8] Alexis Favrot and Christof Faller, "Complementary N-band IIR filterbank based on 2-band complementary filters," in *Proc. of the Intl. Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2010.

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

[1] `http://www.fftw.org/`

[2] `http://ant.hsu-hh.de/upmix`

a) Frequency Domain

	Operations per block			
	ADD	MULT	SQRT	Overall ($2 \rightarrow C$)
STFT				$2 \cdot 5\ \log_2(N_F) \cdot N_F$
Inv. STFT				$C \cdot 5\ \log_2(N_F) \cdot N_F$
Pos. Estimation	$5 \cdot N_F$	$7 \cdot N_F$	$3 \cdot N_F$	$15 \cdot N_F$
Dir./Amb. Separation	$5 \cdot N_F$	$8 \cdot N_F$		$13 \cdot N_F$
Repanning	$(2\,C_1 + 4) \cdot N_F$	$(2\,C_1 + 9) \cdot N_F$	$2 \cdot N_F$	$(4\,C_1 + 15) \cdot N_F$
Overall per block				$\left[\ C \cdot 5\ \log_2(N_F)\ +\ 4\,C_1\ +\ 10\ \log_2(N_F) + 43\ \right] \cdot N_F$
Overall per sample				$\left[\ C \cdot 5\ \log_2(N_F)\ +\ 4\,C_1\ +\ 10\ \log_2(N_F) + 43\ \right] \cdot 2$
- Sample rate 44.1 kHz, $N_F = 2048$				13.5 MFlops $+$ $\left[C \cdot 4.85 + C_1 \cdot 0.35\right]$ MFlops
- Transforms + Upmix				$\left[(C+2) \cdot 4.85\right]$ MFlops $+$ $\left[3.8 + C_1 \cdot 0.35\right]$ MFlops
- $C = 5, C_1 = 3$				34.0 MFlops $+$ 4.85 MFlops

b) Time Domain

	Operations per sample and band			
	ADD	MULT	SQRT	Overall ($2 \rightarrow C$)
Filterbank				$2 \cdot \left[\ M \cdot (3\,N + 4)\ +\ M_A \cdot 3\,N_1\ \right]$
Power estimation	2	6		$8 \cdot (M-1)$
Synthesis	C			$C \cdot (M-1)$
Pos. Estimation	3	3	3	$9 \cdot (M-1)$
Dir./Amb. Separation	3	4		$7 \cdot (M-1)$
Repanning	$(C_1 + 4)$	$(C_1 + 9)$	2	$(2\,C_1 + 15) \cdot (M-1)$
Overall at sample rate 44.1 kHz, $C = 5, C_1 = 3$				
- FB I + Upmix				5.1 MFlops $+$ 3.3 MFlops
- FB II + Upmix				9.6 MFlops $+$ 6.5 MFlops
- FB III + Upmix				13.8 MFlops $+$ 6.5 MFlops

Table 4: *Required operations for an upmix from* 2 *to* C *channels where direct signal repanning is limited to a subset of* C_1 *loudspeakers (e.g. front loudspeakers only).*

[10] 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.

[11] Tapani Pihlajamäki, Olli Santala, and Ville Pulkki, "Synthesis of spatially extended virtual sources with time-frequency decomposition of mono signals," *Journal of the Audio Engineering Society*, vol. 62, no. 7-8, pp. 467–484, 2014.

[12] 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.

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

[14] Alexis Favrot and Christof Faller, "Designing sets of N doubly complementary IIR filter," in *Proc. of the 130th AES Convention*, 2011.

[15] Sanjit K. Mitra, Yrjö Neuvo, and Palghat P. Vaidyanathan, "Complementary IIR digital filter banks," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1985.

[16] Palghat P. Vaidyanathan, Sanjit K. Mitra, and Yrjö Neuvo, "A new approach to the realization of low-sensitivity IIR digital filters," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 2, pp. 350–361, 1986.

[17] Lajos Gazsi, "Explicit formulas for lattice wave digital filters," *IEEE Transactions on Circuits and Systems*, vol. 32, no. 1, pp. 68–88, 1985.

[18] 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.

[19] Franz Zotter, Matthias Frank, Georgios Marentakis, and Alois Sontacchi, "Phantom source widening with deterministic frequency dependent time delays," in *Proc. of the 14th Int. Conference on Digital Audio Effects*, 2011.