DOWNMMIX-COMPATIBLE CONVERSION FROM MONO TO STEREO IN TIME- AND FREQUENCY-DOMAIN

Marco Fink, Sebastian Kraft, Udo Zölzer

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

ABSTRACT

Even in a time of surround and 3D sound, many tracks and recordings are still only available in mono or it is not feasible to record a source with multiple microphones for several reasons. In these cases, a pseudo stereo conversion of mono signals can be a useful preprocessing step and/or an enhancing audio effect. The conversion proposed in this paper is designed to deliver a neutral sounding stereo image by avoiding timbral coloration or reverberation. Additionally, the resulting stereo signal is downmix-compatible and allows to revert to the original mono signal by a simple summation of the left and right channels. Several configuration parameters are shown to control the stereo panorama. The algorithm can be implemented in time-domain or also in the frequency-domain with additional features, like center focusing.

1. INTRODUCTION

A noteworthy amount of recordings was and is still being done in mono for technical or pragmatic reasons. For example, basic broadcast program outside the studio environment is often recorded using a single microphone. Nevertheless, the later replay and mixing would benefit from a stereo recording in terms of spaciousness and pleasing sound. Therefore, the process to create stereo signals from mono recordings, also called pseudo-stereo, is a well-known and broadly utilized field of audio technology. While real stereo mixes are created by panning discrete sources to a specific position in the stereo panorama, pseudo-stereo does usually not involve knowledge about the sources. It rather randomly pans certain frequency components to the left and the right to achieve a decorrelation between both channels.

Early attempts to realize a mono-to-stereo (M2S) conversion used a delayed version of the input signal to provide a second channel [1]. The same author came up with the idea of applying complementary comb filters which were later extended to be phase-aligned in [2]. Alternatively, in [3, 4] different allpass network designs were proposed to obtain a strong decorrelation and to achieve a wide and also scalable stereo image. Another extension allowing more control is shown in [5]. For even stronger decorrelation, a frequency-domain filter design method is suggested in [6]. The above methods either impose a strong timbral coloration or they are not downmix compatible and reversible. Both are important features, though.

A completely different approach granting possibilities to design a specific auditory image is explained in [7]. However, it is not a pseudo stereo algorithm in the narrower sense as it requires complex user input to explicitly define pan positions for certain frequency bands and does not allow a fully automatic conversion. The same holds true for upmixing based on Directional Audio Coding (DirAC) [8]. However, the decorrelation mechanism in the DirAC synthesis are similar to the proposed approach to a certain extent.

The novel pseudo-stereo conversion algorithm is derived in Sec. 2 and its implementation in time- and frequency-domain are depicted in Sec. 3 and Sec. 4, respectively. Section 5 demonstrates the methods capabilities with a few experiments before concluding thoughts are provided in Sec. 6.

2. CONVERSION APPROACH

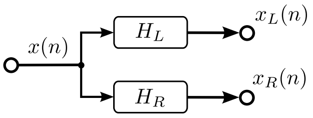


Figure 1: *Basic blockscheme of the pseudo-stereo system*

The basic idea of the pseudo-stereo system is to apply a filter $H_L(e^{j\Omega})$ to an input signal $x(n)$ to produce a first output channel $x_L(n)$ and a different filter $H_R(z)$ in parallel, producing a second output channel $x_R(n)$ as shown in Fig. 1. The static filters should have quite different characteristics to achieve a strong decorrelation between the two output channels and hence, introduce an impression of spatial width.

One major design criterion of the proposed system is downmix compatibility

$$x_M(n) = x_L(n) + x_R(n) = \\ = x(n) * h_L(n) + x(n) * h_R(n) \overset{!}{=} x(n-D), \quad (1)$$

where the downmix signal $x_M(n)$ as the sum of the left and right stereo channel is identical to the unprocessed input except for a delay D. Transferring Eq. (1) to the frequency domain leads to

$$X_M(e^{j\Omega}) = X(e^{j\Omega}) \cdot H_L(e^{j\Omega}) + X(e^{j\Omega}) \cdot H_R(e^{j\Omega}) = \\ = X(e^{j\Omega}) \cdot \Big(H_L(e^{j\Omega}) + H_R(e^{j\Omega})\Big) = \\ \overset{!}{=} X(e^{j\Omega}) \cdot e^{-j\Omega D} \quad (2)$$

and directly imposes the constraint

$$H_L(e^{j\Omega}) + H_R(e^{j\Omega}) \overset{!}{=} e^{-j\Omega D}, \quad (3)$$

where the sum of both transfer functions has constant magnitude and linear phase to guarantee downmix compatibility. Furthermore, for a neutral sounding stereo output without coloration artifacts the sum of the magnitude frequency responses

$$|H_L(e^{j\Omega})| + |H_R(e^{j\Omega})| \overset{!}{=} 1 \tag{4}$$

has to be constant. Additionally, both pseudo-stereo filters must feature conjugate symmetry

$$H_{L/R}(e^{j\Omega}) = H^*_{L/R}(e^{-j\Omega}) \tag{5}$$

to obtain real-valued impulse responses. Since these filters are linear phase with a constant group delay, they will not introduce a phase shift between the stereo output channels and only generate amplitude differences. Therefore, the resulting pseudo-stereo effect is only based on amplitude panning and not on time-delay panning. Hence, $|H_{L/R}(e^{j\Omega})|$ can be interpreted as panning coefficient in the range of $[0,\ldots,1]$ corresponding to full panning from the complementary to the current channel at a certain frequency Ω. A value of 0.5 indicates center panning.

It was found that a regular pattern in the frequency domain, like provided by higher-order complementary comb filters, achieve great decorrelation but due to the uniform frequency response sound very unnatural. Hence, the frequency response of the system was designed to be diffuse and unstructured as shown in Fig. 2.

The actual implementation and the following filter design is achieved in the discrete fourier domain. To represent the frequency indexes of the sampled spectrum the variable k is used in the following. The proposed pseudo stereo filter design is based on a discrete real-valued noise sequence $R(k)$. The noise has a gaussian distribution with a standard deviation σ and a mean value of 0. The sequence $R(k)$ is scaled with w^2 and non-linearly clamped with an $\arctan$ function. Further normalization by π and an additive offset of $\frac{1}{2}$ yields the transfer function

$$H_L(k) = \left(\frac{1}{2} + \frac{1}{\pi}\arctan(w^2 \cdot R(k))\right) e^{-j\frac{2\pi k D}{N}} \tag{6}$$

of the left channel decorrelation filter. The amplitudes of that and its complementary filter $H_R(k) = 1 - H_L(k)$ are bound to a range $[0,\ldots,1]$ and fulfill the requirements defined in (3) and (4). The parameter w allows dynamic control of the resulting stereo width. For $w = 0$ the filters have a constant magnitude response of 0.5 and hence, no panning is performed. With increasing w the values of the frequency responses are more and more clamped to the extreme values 0 and 1 (Fig. 2). In consequence, a higher degree of decorrelation is achieved.

The panning of very low and very high frequencies can be perceived annoying since the listener is used to hear certain instruments like bass guitar and bass drum in the center of the stereo panorama. Therefore, the amplitude panning in those frequency regions is disabled by setting

$$|H_{L/R}(k)| = 0.5, \quad \text{for}\, k < k_{\text{lo}} \vee k > k_{\text{hi}}, \tag{7}$$

where $k_{\text{lo/hi}}$ are the cut-off frequencies defining the passband to be actually processed.

An exemplary M2S conversion is shown in Fig. 4. A recording of a small singing ensemble with 4 voices is used as input signal $x(n)$. The spectrogram in Fig. 4a) shows the trend of the harmonic voices. The spectrograms for left and right output channel in Fig. 4b+c) indicate how the input signal was distributed. The alto voice at about 1000 Hz is mainly panned to the left channel, whereas the fundamental of the soprano voice at about 2200 Hz can be found in the right channel. The lower passband cutoff is set to $f_{\text{lo}} = 300$ Hz in this example. This results in the center panning of the bass voice showing a fundamental frequency of about 150 Hz.

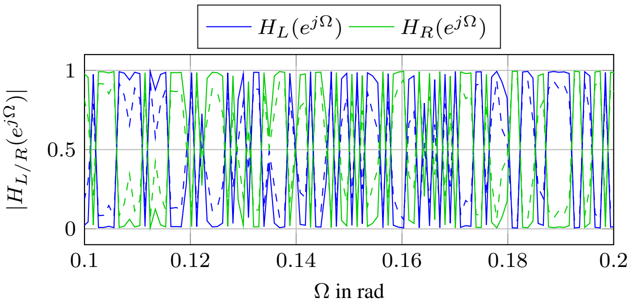


Figure 2: *Exemplary frequency response of the decorrelation filters for standard deviation* $\sigma = 25$, *stereo width* $w = 1$ *(solid) and* $w = 0.1$ *(dashed)*

3. TIME-DOMAIN IMPLEMENTATION

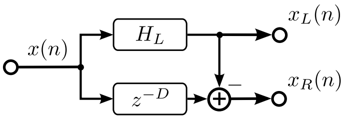


Figure 3: *Blockscheme of the time-domain realization*

The pseudo-stereo conversion can be performed in time-domain by FIR filtering the input signal with the impulse response

$$h_L(n) = \mathcal{F}^{-1}\{H_L(k)\} \tag{8}$$

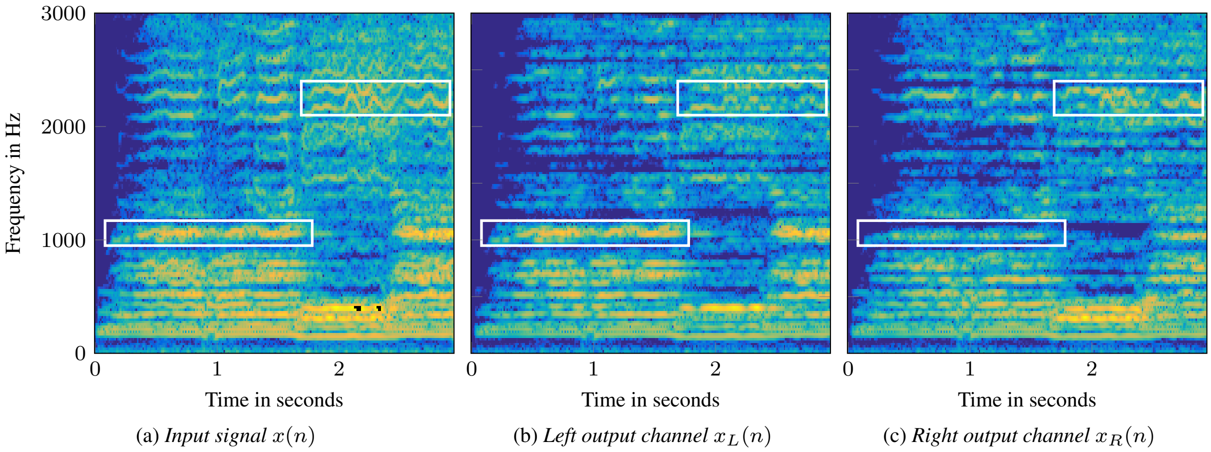
to obtain the left channel. Due to (3) and (4), the right channel can be computed by simply subtracting the FIR filter output from the time-delayed input signal

$$x_R(n) = x(n-D) - x_L(n), \tag{9}$$

where $D = \frac{N-1}{2}$ is the group delay of the FIR filter of length N. This only requires a single FIR filter together with a delay line and hence, is easy to realize on various platforms. On the other hand, the filter is static and the filtering operation for long impulse responses is computationally expensive. Furthermore, no dynamic control on the actual panning is provided and for example strong sources that are expected in the center of the stereo panorama could be diffusely panned. As the filter design is already done in the frequency domain it is a logical consequence to make use of fast convolution to reduce the complexity of the filtering and at the same time use the spectral information from the input signal do derive guidelines for the filter design process.

4. FREQUENCY-DOMAIN IMPLEMENTATION

First of all, the input signal $x(n)$ has to be transferred to the time-frequency domain with the help of a short-time fourier transform



(a) *Input signal $x(n)$* (b) *Left output channel $x_L(n)$* (c) *Right output channel $x_R(n)$*

Figure 4: *Exemplary M2S conversion of a choir sample*

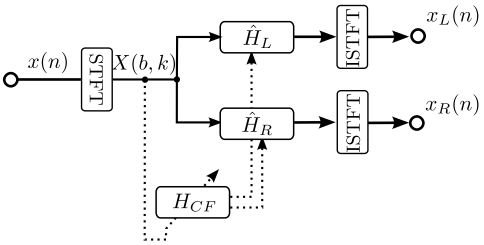


Figure 5: *Blockscheme of the frequency-domain realization*

(STFT) yielding the spectra $X(b,k)$, where b and k denote the block index and frequency index. The time-frequency representations of the output channels

$$X_{L/R}(b,k) = X(b,k) \cdot H_{L/R}(k) \tag{10}$$

are now computed with a dot-wise product and a following overlapping inverse STFT synthesis of the left and right output channel.

Until now, the pseudo stereo filters were applied statically. Hence, all sound sources of the mono input signal were diffusely panned to create a wide stereo panorama to enrich the input signal with a spacious character. Nevertheless, the diffuse panning of a dominant source like the singing voice in a musical piece constitutes a very unusual, discomforting listening experience. Therefore, the authors extended the M2S conversion scheme with a so-called center-focusing filter $H_{CF}(b,k)$ which aims to keep dominant sources in the center of the stereo panorama. Two features were considered to assess the dominance of a sound source and to compute $H_{CF}(b,k)$.

4.1. Normalized Energy Estimate

Assuming that signal sources expected to be in the center of the stereo mix feature the largest spectral energy, the center-focusing filter is defined as

$$H_{CF}(b,k) = (1-\alpha)\, H_{CF}(b-1,k) + \alpha\, X_n^2(b,k) \tag{11}$$

$$X_n(b,k) = \frac{|X(b,k)|}{\max\limits_k |X(b,k)|} \tag{12}$$

in the first variant. For every frame b, an amplitude-normalized spectrum $X_n(b,k)$ is computed using the maximum value of the current magnitude spectrum $|X(b,k)|$. $X_n(b,k)$ is then squared, weighted by α and added to $(1-\alpha)\,H_{CF}(b,k)$ to obtain a recursively smoothed estimation of the high energy regions in the spectrum.

4.2. Tonality

The second variant is based on tonalness measures. Several features to identify the tonalness of a signal were presented in [9]. In this study, the tonalness from the amplitude threshold feature $t_{\text{AT}}(b,k)$ and peakiness features $t_{\text{PK}}(b,k)$ are combined, resulting in the center-focusing filter

$$H_{CF}(b,k) = t_{\text{AT}}(b,k) \cdot t_{\text{PT}}(b,k). \tag{13}$$

4.3. Application of center-focusing filter

To involve the center-focusing filter, the fixed pseudo stereo filters $H_{L/R}(b,k)$ are weighted accordingly

$$\hat{H}_{L/R}(b,k) = H_{L/R}(k)\, H_{CF}(b,k) - \frac{1}{2}\,(1 - H_{CF}(b,k)) \tag{14}$$

to force the filters magnitude transfer function to a value of 0.5 for high values of $H_{CF}(b,k)$. The weighted pseudo stereo filters $\hat{H}_{L/R}(b,k)$ are computed for every frame b and applied to the input signal $X(b,k)$ as shown in Fig. 5. The center-focusing and the combined filter still meet the conditions of Eq. (1)-(4). The additional complexity, caused by the computation of $\hat{H}_{L/R}(b,k)$, is legitimated by the enriched naturality of the M2S stereo panorama.

5. EXPERIMENTS

A typical measurement to describe channel decorrelation or the width of a stereo panorama, especially for room measurements, is the so-called Interchannel Crosscorrelation Coefficient (ICC)

$$\text{ICC} = \max |\rho_{LR}(\tau)|, \tag{15}$$

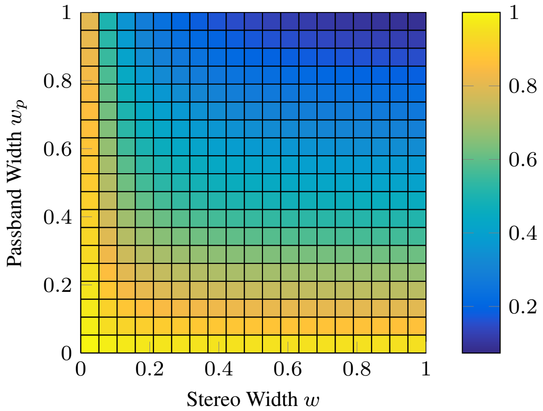


Figure 6: *ICC for white noise input and varying passband width and stereo width*

which is defined as the maximum value of the normalized cross-correlation function

$$\rho_{LR}(\tau) = \frac{\sum x_L(n)x_R(n-\tau)}{\sum x_L^2(n)\sum x_R^2(n-\tau)}. \qquad (16)$$

The ICC for a white noise mono input after processing with the filter design from Eq. (6) and (7) in dependency of stereo width w and relative passband width $w_p = 1 - 2f_{\mathrm{lo}}/f_s$ is plotted in Fig. 6. Fully correlated signals, resulting in ICC $= 1$, are obtained for $w = 0$ and $w_p = 0$. For increasing w and w_p the ICC value is decreasing continuously. The smallest value of 0.0637 indicates an almost full decorrelation for the maximum values of w and w_p. A FIR filter was designed using Eq. 6 with a length of $N = 2048$ samples, standard deviation $\sigma = 25$, and a sampling rate of $f_s = 44100$ in this example.

Another tool of audio engineers to graphically judge a stereo mix is the audio goniometer. It is basically a X-Y illustration of a stereo signal as shown in Fig. 7 and was originally created with oscilloscopes. A very narrow stereo mix or a mono signal would be illustrated as a straight line, whereas balanced stereo signals featuring level and phase differences tend to appear sphere-like. The stereo signal used for this example was a short pop music excerpt. The sphere-like shape of the original signal can be easily identified in Fig. 7. The downmix that is fed to the M2S system appears as a straight line whereas the M2S converted signal again features a sphere-like shape with a similar width as the original indicating a comparable stereo width.

To allow interested readers to experience the benefit of the M2S conversion, the authors provide some rendered wavefiles that can be found at `http://ant.hsu-hh.de/dafx15/M2S`.

6. CONCLUSION

This study proposed an approach to convert mono into stereo signals using pure amplitude panning. Necessary filter design constraints to perform downmix-compatible pseudo-stereo conversion without any timbral coloration were derived. The versatile design offers various control parameters to adjust stereo width and the cut-off frequencies of the passband to be processed. Subsequently, the described conversion system is implemented in time- and frequency-domain.

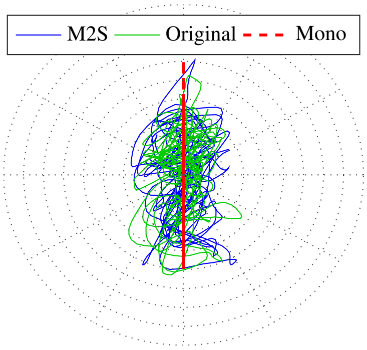


Figure 7: *Goniometer of a music sequence showing a reference stereo track, its downmix, and the M2S converted track*

The potential problem of a convenient listening experience due to a missing center focus is solved by extending the system with a center-focusing filter. This filter selects dominant spectral components according to predefined features, like normalized energy or tonalness, and repans them to the center.

Measurements proved the possibility of achieving nearly complete decorrelation, corresponding to extreme panning, when the conversion system is operated with extreme settings. For moderate settings, the mono-to-stereo conversion produces pleasing and natural sounding stereo panoramas that clearly enhance mono recordings with the arising spaciousness.

7. REFERENCES

[1] Holger Lauridsen, “Nogle forsog reed forskellige former rum akustik gengivelse,” *Ingenioren*, vol. 47, pp. 906, 1954.

[2] Manfred R. Schroeder, “An artificial stereophonic effect obtained from a single audio signal,” *J. Audio Eng. Soc*, vol. 6, no. 2, pp. 74–79, 1958.

[3] Benjamin B. Bauer, “Some techniques toward better stereophonic perspective,” *J. Audio Eng. Soc*, vol. 17, no. 4, pp. 410–415, 1969.

[4] Robert Orban, “A rational technique for synthesizing pseudo-stereo from monophonic sources,” *J. Audio Eng. Soc*, vol. 18, no. 2, pp. 157–164, 1970.

[5] Michael A. Gerzon, “Signal processing for simulating realistic stereo images,” in *Audio Engineering Society Convention 93*, Oct 1992.

[6] Gary S. Kendall, “The decorrelation of audio signals and its impact on spatial imagery,” *Computer Music J.*, vol. 19, no. 4, pp. 72–87, 1995.

[7] Christof Faller, “Pseudostereophony revisited,” in *Audio Engineering Society Convention 118*, May 2005.

[8] Archontis Politis, Tapani Pihlajamäki, and Ville Pulkki, “Parametric Spatial Audio Effects,” *Proc. of the 15th Int. Conference on Digital Audio Effects (DAFx-12)*, 2012.

[9] Sebastian Kraft, Alexander Lerch, and Udo Zölzer, “The Tonalness Spectrum: Feature-Based Estimation of Tonal Components,” *Proc. of the 16th Int. Conference on Digital Audio Effects (DAFx-13)*, 2013.